Electronic Colloquium on Computational Complexity, Report No. 176 (2023)

A Technique for Hardness Amplification Against AC^0

William M. Hoza[*]
Department of Computer Science
The University of Chicago
williamhoza@uchicago.edu

Abstract

We study hardness amplification in the context of two well-known "moderate" average-case hardness results for AC^0 circuits. First, we investigate the extent to which AC^0 circuits of depth d can approximate AC^0 circuits of some larger depth $d+k$. The case $k=1$ is resolved by Håstad, Rossman, Servedio, and Tan's celebrated average-case depth hierarchy theorem (JACM 2017). Our contribution is a significantly stronger correlation bound when $k \geq 3$. Specifically, we show that there exists a linear-size AC^0_{d+k} circuit $h\colon \{0,1\}^n \to \{0,1\}$ such that for every AC^0_d circuit g, either g has size $\exp(n^{\Omega(1/d)})$, or else g agrees with h on at most a $(1/2+\varepsilon)$-fraction of inputs where $\varepsilon = \exp(-(1/d) \cdot \Omega(\log n)^{k-1})$. For comparison, Håstad, Rossman, Servedio, and Tan's result has $\varepsilon = n^{-\Theta(1/d)}$. Second, we consider the majority function. It is well known that the majority function is moderately hard for AC^0 circuits (and stronger classes). Our contribution is a stronger correlation bound for the XOR of t copies of the n-bit majority function, denoted $\mathsf{MAJ}_n^{\oplus t}$. We show that if g is an AC^0_d circuit of size S, then g agrees with $\mathsf{MAJ}_n^{\oplus t}$ on at most a $(1/2+\varepsilon)$-fraction of inputs, where $\varepsilon = \left(n^{-1/2} \cdot O(\log S)^{d-1} \cdot \sqrt{\log n}\right)^t$.

To prove these results, we develop a hardness amplification technique that is tailored to a specific type of circuit lower bound proof. In particular, one way to show that a function h is moderately hard for AC^0 circuits is to (a) design some distribution over random restrictions or random projections, (b) show that AC^0 circuits simplify to shallow decision trees under these restrictions/projections, and finally (c) show that after applying the restriction/projection, h is moderately hard for shallow decision trees with respect to an appropriate distribution. We show that if h can be proven to be moderately hard by a proof with that structure, then XORing multiple copies of h amplifies its hardness. Our analysis involves a new kind of XOR lemma for decision trees, which might be of independent interest.

1 Introduction

1.1 Average-Case Circuit Lower Bounds

Circuit lower bounds are at the heart of computational complexity theory. To understand the limitations of (extremely) efficient computation, we seek to prove that certain explicit functions cannot be computed by certain interesting classes of Boolean circuits. In fact, ideally, we want to prove *average-case* circuit lower bounds, also known as *correlation bounds*. That is, we would like to prove that circuits in some class $\mathcal{C}$ cannot compute some function $h\colon \{0,1\}^n \to \{0,1\}$ on more than a $(1/2+\varepsilon)$-fraction of inputs for some small value $\varepsilon > 0$:

$$\text{For every } g \in \mathcal{C}, \qquad \Pr_{\mathbf{x} \in \{0,1\}^n}[g(\mathbf{x}) = h(\mathbf{x})] \leq \frac{1}{2} + \varepsilon. \tag{1}$$

[*] Part of this work was done while the author was visiting the Simons Institute for the Theory of Computing.

ISSN 1433-8092

We would like ε to be as small as possible. For example, one motivation for trying to minimize ε comes from the Nisan-Wigderson framework for converting correlation bounds into pseudorandom generators (PRGs) [NW94]. In this framework, a bound of the form (1) implies a PRG with error εn, and in particular, the framework requires $\varepsilon < 1/n$.

In this work, we focus on the case that $\mathcal{C}$ consists of AC^0 circuits, i.e., circuits made up of AND and OR gates with unbounded fan-in, with literals and constants at the bottom. The *size* of the circuit is the number of AND and OR gates, and the *depth* of the circuit is the length of the longest path from the output gate to an input gate. We refer to an AC^0 circuit of depth d as an "AC^0_d circuit." We are especially interested in the constant-depth regime; this class of circuits can be viewed as a model of constant-time parallel computation. Some of the most celebrated theorems in circuit complexity are lower bounds on the size of AC^0 circuits computing various explicit functions. For example, if g is an AC^0_d circuit, then g famously cannot compute the parity function on n bits or the majority function on n bits, unless g has size at least $\exp(c_d \cdot n^{1/(d-1)})$ [FSS84; Ajt83; Yao85; Hås86a; Hås86b].

1.2 Hardness Amplification and Yao's XOR Lemma

One appealing approach for proving strong correlation bounds is to first construct a function h that is "moderately hard" (e.g., maybe we have $\varepsilon = 1/\sqrt{n}$), and then apply some kind of *hardness amplification* scheme that converts h into a "very hard" function (e.g., maybe now we can take $\varepsilon = n^{-\omega(1)}$). The most famous method for hardness amplification is Yao's XOR Lemma [Yao82; Lev87; Imp95; GNW11]. Starting from a hard function $h\colon \{0,1\}^n \to \{0,1\}$, this lemma considers the new hard function $h^{\oplus t}\colon \{0,1\}^{nt} \to \{0,1\}$ defined by $h^{\oplus t}(x^{(1)}, \dots, x^{(t)}) = \bigoplus_{i=1}^{t} h(x^{(i)})$. One well-known version[1] of Yao's XOR Lemma says that if h is moderately hard for $\mathsf{MAJ} \circ \mathcal{C}$ circuits, where MAJ denotes the majority function, then $h^{\oplus t}$ is very hard for $\mathcal{C}$ circuits.

In the context of relatively weak classes such as AC^0, the distinction between $\mathcal{C}$ and $\mathsf{MAJ} \circ \mathcal{C}$ is extremely important. Proving lower bounds on the size of $\mathsf{MAJ} \circ \mathcal{C}$ circuits is generally much more difficult than proving lower bounds on the size of $\mathcal{C}$ circuits. For this reason, there is a great deal of interest in "removing the majority gate" from Yao's XOR Lemma. For example, we can ask the following.

Question 1 (Does XORing amplify hardness for AC^0?)**.** *Let $h\colon \{0,1\}^n \to \{0,1\}$ and let $t = \log n$. Assume that every constant-depth subexponential-size AC^0 circuit g satisfies*

$$\Pr_{\mathbf{x} \in \{0,1\}^n}[g(\mathbf{x}) = h(\mathbf{x})] \le \frac{1}{2} + n^{-\Omega(1)}.$$

Does it follow that every constant-depth polynomial-size AC^0 circuit g satisfies

$$\Pr_{\mathbf{x} \in \{0,1\}^{nt}}[g(\mathbf{x}) = h^{\oplus t}(\mathbf{x})] \le \frac{1}{2} + n^{-\omega(1)}\,?$$

Several recent papers have developed and applied a refined version of Yao's XOR Lemma featuring an "approximate linear sum" gate instead of the traditional majority gate [CLW20; CL21; CLLO21; HV21; Che23; CHLR23]. This clever approach has been fruitful, but it is still not applicable if we start with a function that is hard merely for AC^0 circuits. Unfortunately, there are strong *barrier results* saying that every "black-box" hardness amplification scheme must involve *some* nontrivial computational overhead [Vio06; GR08; SV10; GSV18; Sha23]. As a special case,

[1] See, for example, Viola's work [Vio20].

this line of work implies that Question 1 cannot be resolved affirmatively via a "black-box" hardness amplification scheme. Thus, we have an ironic state of affairs: we have a rich toolkit for proving lower bounds on the size of AC^0 circuits, because we are able to exploit these circuits' weaknesses, but at the same time, *specifically because these circuits are too weak*, we cannot use Yao's XOR Lemma to amplify our lower bounds.[2]

1.3 Our Contributions

In this work, we develop a non-black-box method for hardness amplification, applicable to some (but not all) moderate hardness results for AC^0 circuits. We use our method to amplify two well-known average-case hardness results, discussed next.

1.3.1 Correlation Bounds for Depth Reduction Within AC^0

Our first application of our hardness amplification technique concerns the role of depth in circuit complexity. To what extent are deeper circuits intrinsically more powerful than shallower circuits? In other words, what is the *marginal utility of time* for parallel computation?

Surprisingly, it turns out that in many contexts, circuits can be generically and nontrivially simulated by shallower circuits. For example:

- NC^1 circuits (i.e., circuits of depth $O(\log n)$ with bounded fan-in) can be simulated by AC^0_d circuits of size $\exp(n^{O(1/d)})$ [Val77; Vio09; Vio17; Tel20].
- ACC^0_d circuits (i.e., AC^0_d circuits augmented with MOD_m gates) of size S can be simulated by $\mathsf{SYM} \circ \mathsf{AND}$ circuits of size $\exp((\log S)^{O(d)})$ [Tod91; All89; AH94; Yao90; AG94; BT94; Wil14; CP19].
- AC^0 circuits can be approximated in various ways by low-degree polynomials [Raz87; Smo87; Smo93; BRS91; Tar93; LMN93; Bop97; Hås01; Baz09; Raz09; Bra10; Tal17; KS18; HS19], which can be viewed as a "depth-two" model of computation.

In light of these remarkable "depth reduction" results and their numerous applications, we would like to know precisely when, and to what extent, depth reduction is possible. Indeed, there is a longstanding interest in thoroughly understanding the *hardness of circuit depth reduction* within AC^0. Early work shows that there exists a linear-size AC^0_{d+1} circuit $h\colon \{0,1\}^n \to \{0,1\}$ such that every AC^0_d circuit computing h must have size $\exp(n^{\Omega(1/d)})$ [Sip83; Yao85; Hås86a]. For several decades, it was a stubborn open problem to prove a similar hierarchy theorem in the average-case setting. O'Donnell and Wimmer essentially resolved the depth-2 vs. depth-3 case [OW07], and then finally Håstad, Rossman, Servedio, and Tan resolved the general depth-d vs. depth-$(d+1)$ case in a breakthrough last decade [HRST17]:

Theorem 1 (The average-case depth hierarchy theorem [HRST17])**.** *Let $n, d \in \mathbb{N}$ with $d \leq \frac{\alpha \log n}{\log \log n}$, where $\alpha > 0$ is a suitable constant. There is an explicit*[3] AC^0_{d+1} *circuit $h\colon \{0,1\}^n \to \{0,1\}$ of size $O(n)$ such that for every* AC^0_d *circuit $g\colon \{0,1\}^n \to \{0,1\}$, either g has size $\exp(n^{\Omega(1/d)})$, or else the following correlation bound holds:*

$$\Pr_{\mathbf{x} \in \{0,1\}^n}[g(\mathbf{x}) = h(\mathbf{x})] \leq \frac{1}{2} + n^{-\Omega(1/d)}. \tag{2}$$

[2] The exception, of course, is if we start from a lower bound against a stronger class such as $\mathsf{MAJ} \circ \mathsf{AC}^0$. See Klivans' work [Kli01].

[3] I.e., the circuit h can be constructed in $\mathrm{poly}(n)$ time, given the parameters n and d.

Theorem 1 asserts that h is *moderately* hard for AC^0_d circuits. Håstad, Rossman, Servedio, and Tan identified two obstacles preventing significant improvement of the $n^{\Omega(1/d)}$ correlation bound in (2):

- The "hard function" h in Theorem 1 is monotone. By the Kahn-Kalai-Linial theorem [KKL88], every monotone Boolean function can be approximated by a constant or a variable with success probability $1/2 + \omega(1/n)$.
- By the discriminator lemma [HMPST93], every linear-size AC^0_{d+1} circuit h, whether monotone or not, can be approximated by a linear-size AC^0_d circuit with success probability $1/2 + \Omega(1/n)$.

(See Hatami, Hoza, Tal, and Tell's work for further details of these two arguments [HHTT23, Appendix A].)

In this work, we overcome both obstacles by using a different, non-monotone hard function h with depth slightly greater than $d + 1$. We prove an average-case lower bound for the task of simulating AC^0_{d+k} circuits using AC^0_d circuits, with a correlation bound that gets significantly stronger as k gets larger.

Theorem 2 (AC^0_d circuits cannot approximate AC^0_{d+k} circuits)**.** *Let $n, d, k \in \mathbb{N}$ with $k \geq 3$ and $dk \leq \frac{\alpha \log n}{\log \log n}$, where $\alpha > 0$ is a suitable constant. There is an explicit AC^0_{d+k} circuit $h\colon \{0,1\}^n \to \{0,1\}$ of size $O(n)$ such that for every AC^0_d circuit $g\colon \{0,1\}^n \to \{0,1\}$, either g has size $\exp(n^{\Omega(1/d)})$, or else the following correlation bound holds:*

$$\Pr_{\mathbf{x} \in \{0,1\}^n}[g(\mathbf{x}) = h(\mathbf{x})] \leq \frac{1}{2} + \exp\left(-\frac{1}{d} \cdot \Omega(\log n)^{k-1}\right).$$

Our hard function h is the XOR of approximately $\log^{k-2} n$ many copies of Håstad, Rossman, Servedio, and Tan's hard function [HRST17]. By combining Theorem 2 with the Nisan-Wigderson framework [NW94] and a reduction due to Li and Zuckerman [LZ19], we obtain new constructions of *seedless randomness extractors* that are computable by small $\mathsf{AC}^0_{d+O(1)}$ circuits and that can extract from sources that are "recognizable" by large AC^0_d circuits. See subsection 4.4 for details.

1.3.2 Correlation Bounds for XOR of Majority

Our second application of our hardness amplification technique concerns the n-bit majority function (MAJ_n). It is well known that the majority function is moderately hard for AC^0 circuits and more generally for $\mathsf{AC}^0[\oplus]$ circuits, i.e., AC^0 circuits augmented with parity gates.[4] Specifically, based on the seminal works of Razborov and Smolensky [Raz87; Smo87; Smo93], we have the following correlation bound.

Theorem 3 (Majority is moderately hard for $\mathsf{AC}^0_d[\oplus]$ circuits)**.** *Let $n, d, S \in \mathbb{N}$ with $S \geq n$. Let $g\colon \{0,1\}^n \to \{0,1\}$ be an $\mathsf{AC}^0_d[\oplus]$ circuit of size S. Then*

$$\Pr_{\mathbf{x} \in \{0,1\}^n}[g(\mathbf{x}) = \mathsf{MAJ}_n(\mathbf{x})] \leq \frac{1}{2} + \frac{O(\log S)^{d-1}}{\sqrt{n}}.$$

We emphasize that we are considering the problem of computing the majority function on a $(1/2 + \varepsilon)$-fraction of n-bit inputs, which is distinct from the perhaps more famous "promise majority"

[4] Even more generally, we can consider MOD_q gates where q is a power of a prime – but let us focus on parity gates for simplicity.

problem in which we wish to compute the majority function on all inputs with relative Hamming weight outside the interval $1/2 \pm \varepsilon$. It seems that O'Donnell and Wimmer were the first to explicitly consider correlation bounds for the majority function [OW07].

The specific quantitative bound in Theorem 3 is actually a log-factor improvement over what was known before, to the best of our knowledge. We therefore include a proof of Theorem 3 in Appendix A. (We also present a matching AC^0 construction based on prior work, showing that Theorem 3 is tight.) That being said, our main focus is on the qualitative distinction between functions that are "moderately hard" and functions that are "very hard." The fact that the majority function is moderately hard for $\mathsf{AC}^0[\oplus]$ circuits – for example, the correlation bound above is $\widetilde{O}(\sqrt{n})$ in the constant-depth polynomial-size regime – was already well-understood prior to this work.

Remarkably, this weak correlation bound is the best bound known on the correlation between $\mathsf{AC}^0[\oplus]$ circuits and any hard function in NP.[5] It is a major open problem to construct an explicit function that is provably "very hard" for $\mathsf{AC}^0[\oplus]$ circuits. The function $\mathsf{MAJ}_n^{\oplus t}$, perhaps with $t = \operatorname{polylog}(n)$, seems like a reasonable candidate.

Chattopadhyay, Hatami, Hosseini, Lovett, and Zuckerman recently proved that XORing amplifies the hardness of MAJ_n for constant-degree $\mathbb{F}_2$-polynomials [CHHLZ20], which can be considered a special case of polynomial-size $\mathsf{AC}_2^0[\oplus]$ circuits. In this work, we consider a different special case of $\mathsf{AC}^0[\oplus]$ circuits, namely AC^0 circuits. Our contribution is a proof that XORing amplifies the hardness of MAJ_n for AC^0 circuits, albeit with an extra factor of $\sqrt{\log n}$.

Theorem 4 ($\mathsf{MAJ}_n^{\oplus t}$ is hard for AC_d^0 circuits)**.** *Let $n, t, d, S \in \mathbb{N}$ and let $g\colon \{0,1\}^{nt} \to \{0,1\}$ be an AC_d^0 circuit of size S. Then*

$$\Pr_{\mathbf{x} \in \{0,1\}^{nt}} \left[g(\mathbf{x}) = \mathsf{MAJ}_n^{\oplus t}(\mathbf{x}) \right] \le \frac{1}{2} + \left(\frac{O(\log S)^{d-1} \cdot \sqrt{\log n}}{\sqrt{n}} \right)^t.$$

1.4 Our Technique

1.4.1 The Random Simplification Method

To prove Theorem 2 and Theorem 4, we develop a hardness amplification technique that is tailored to a specific method for proving moderate hardness. To explain, let us first consider the challenge of amplifying Håstad, Rossman, Servedio, and Tan's average-case depth hierarchy theorem [HRST17]. Håstad, Rossman, Servedio, and Tan's lower bound proof is based on the concept of *random projections*, which generalize traditional random restrictions. To prove that their hard function h is moderately hard for AC_d^0 circuits, Håstad, Rossman, Servedio, and Tan carefully designed a distribution $\mathcal{R}$ over projections and a distribution μ over inputs and showed the following [HRST17].

1. (Completion to the uniform distribution.) For every function $f\colon \{0,1\}^n \to \{0,1\}$, plugging a uniform random $\mathbf{x} \in \{0,1\}^n$ into f is equivalent to first sampling a projection $\boldsymbol{\pi} \sim \mathcal{R}$, then independently sampling an input $\mathbf{y} \sim \mu$, and finally plugging $\mathbf{y}$ into $f|_{\boldsymbol{\pi}}$.

2. (Simplification.) For every AC_d^0 circuit g, either g has size $\exp(n^{\Omega(1/d)})$, or else with high probability over $\boldsymbol{\pi} \sim \mathcal{R}$, the circuit g *simplifies* under $\boldsymbol{\pi}$ in the sense that $g|_{\boldsymbol{\pi}}$ can be computed by a shallow decision tree.

3. (Maintaining structure.) With high probability over $\boldsymbol{\pi} \sim \mathcal{R}$, the hard function h *maintains structure* in the sense that $h|_{\boldsymbol{\pi}}$ is moderately hard for shallow decision trees with respect to μ.

[5] If we permit hard functions that satisfy less stringent explicitness conditions, then better correlation bounds are known against $\mathsf{AC}^0[\oplus]$ and even stronger classes [Vio20; CR22; CLW20; Che23].

Taken together, the three steps above imply that h is moderately hard for AC^0_d circuits with respect to a uniform random input. We call this proof structure the *random simplification method* for proving correlation bounds.

In this work, we prove XOR lemmas which say (roughly) that if the random simplification method shows that some function h is moderately hard for some circuit class $\mathcal{C}$, then $h^{\oplus t}$ is very hard for *that same circuit class* $\mathcal{C}$. (There are some additional technical assumptions; see Lemma 4 and Lemma 8 for details.) The basic idea behind our XOR lemmas is to first prove an XOR lemma for decision trees, and then apply it to $h|_{\boldsymbol{\pi}}$. Let us therefore discuss XOR lemmas for decision trees.

1.4.2 XOR Lemmas for Decision Trees

Let h be a Boolean function, and assume that every depth-D decision tree agrees with h on at most a $(1/2+\varepsilon)$-fraction of inputs.

For the simplest version of our hardness amplification technique (see Lemma 4), we merely use a rather trivial XOR lemma for decision trees that was already more-or-less known. In particular, it turns out that decision trees *of that same depth* D can compute $h^{\oplus t}$ on at most a $(1/2+\varepsilon')$-fraction of inputs, where $\varepsilon' = \frac{1}{2}\cdot(2\varepsilon)^t$. (For example, this is a special case of Shaltiel's analysis of "fair" decision trees [Sha03].) A slight generalization of that simple analysis is already sufficient for proving our correlation bound for depth reduction within AC^0 (Theorem 2).

On the other hand, to get the best parameters in Theorem 4 (on the hardness of $\mathsf{MAJ}_n^{\oplus t}$), it turns out that we need a more sophisticated XOR lemma for decision trees, in which we allow the tree attempting to compute $h^{\oplus t}$ to have depth significantly larger than D.

This problem has been previously studied by Drucker [Dru12]. Focusing on one setting of parameters, Drucker showed that for every constant $\alpha > 0$, there is a value $D' = \Omega(Dt)$ such that trees of depth D' cannot compute $h^{\oplus t}$ on more than a $(1/2+\varepsilon')$-fraction of inputs, where $\varepsilon' = O(\varepsilon)^{(1-\alpha)\cdot t}$ [Dru12]. Although it comes close, this result is not quite sufficient to prove Theorem 4 because of the $(1-\alpha)$-factor loss in the exponent. Furthermore, unfortunately, the $(1-\alpha)$-factor loss is unavoidable in general, due to counterexamples identified by Shaltiel [Sha03]. The idea behind these counterexamples is that although h is hard for decision trees of depth D, it might nevertheless be easy for decision trees of depth $D+1$. In this case, for any constant $c > 0$, a decision tree of depth cDt can successfully compute h on $\Omega(t)$ independent instances.

To circumvent Shaltiel's counterexamples [Sha03], we strengthen the assumption. We assume that h is moderately hard for depth-D decision trees *for all* D *simultaneously*, with a correlation bound ε that scales with the depth D according to some log-concave function $\varepsilon(D)$. Under this assumption, we prove the decision trees of depth $\Omega(Dt)$ have correlation at most $O(\varepsilon)^t$ with $h^{\oplus t}$.

Lemma 1 (XOR lemma for decision trees under a robust hardness assumption)**.** *Let $h\colon \{0,1\}^n \to \{0,1\}$ be a function and let $\varepsilon\colon [0,\infty) \to (0,\infty)$ be a log-concave function. Assume that for every $D \in \mathbb{N}$ and every decision tree $T\colon \{0,1\}^n \to \{0,1\}$ of depth at most D, we have*

$$\Pr_{\mathbf{x}\in\{0,1\}^n}[T(\mathbf{x}) = h(\mathbf{x})] \le \frac{1}{2} + \varepsilon(D).$$

Then for every $D, t \in \mathbb{N}$ and every decision tree $T\colon \{0,1\}^{nt} \to \{0,1\}$ of depth at most $Dt/2$, we have

$$\Pr_{\mathbf{x}\in\{0,1\}^{nt}}[T(\mathbf{x}) = h^{\oplus t}(\mathbf{x})] \le \frac{1}{2} + O(\varepsilon(D))^t.$$

(See Lemma 3 for a more general statement.)

1.4.3 Using Our XOR Lemmas to Prove Our Correlation Bounds

Given the XOR lemmas for decision trees discussed above, our XOR lemmas for the random simplification method readily follow. In turn, these lemmas imply our correlation bound for depth reduction within AC^0 (Theorem 2) by inspection of Håstad, Rossman, Servedio, and Tan's proof [HRST17].

On the other hand, more effort is required for our proof of the hardness of $\mathsf{MAJ}_n^{\oplus t}$ (Theorem 4). There are at least three known proofs that the majority function is moderately hard for AC^0 circuits: one using the Razborov-Smolensky method [Fil10] (see also Appendix A), one due to O'Donnell and Wimmer [OW07], and one due to Tal [Tal17]. The issue is that none of these proofs fits into our framework of "random simplification arguments." (The latter two proofs do use switching lemmas, but only in an indirect Fourier-analytic way.) For this reason, in subsection 5.1, we present yet another proof that the majority function is moderately hard for AC^0_d circuits. Unfortunately, the correlation bound we get is worse than the optimal bound by a factor of $\sqrt{\log n}$, but the important thing is that our proof is a random simplification argument. Furthermore, crucially, the "robust hardness assumption" of Lemma 1 is satisfied in our proof. Therefore, we are able to combine our proof with our XOR lemma to complete our analysis of $\mathsf{MAJ}_n^{\oplus t}$.

1.5 Related Work

Hardness amplification for weak circuit classes. Goldwasser, Gutfreund, Healy, Kaufman, and Rothblum designed a method for converting worst-case hardness into moderate average-case hardness in the context of weak circuit classes [GGHKR07], which complements our work in some ways. One contrast between their work and ours is that they merely construct a hard function with a very weak explicitness guarantee, namely membership in EXP, whereas we study an extremely explicit hardness amplification method, namely XORing. More recently, Chen, Lu, Lyu, and Oliveira developed a method for constructing very hard functions for weak circuit classes starting from relatively weak assumptions [CLLO21] – but once again, their hard functions only satisfy weak explicitness guarantees such as membership in E.

Klivans' proof that parity is average-case-hard for AC^0 circuits. A long sequence of works has established strong bounds on the correlation between the parity function and AC^0 circuits [FSS84; Ajt83; Yao85; Hås86a; Hås86b; Bab87; Kli01; Vio09; BIS12; IMP12; Hås14]. One of these works, by Klivans [Kli01], is especially relevant for us. Klivans' proof is based on a result by Aspnes, Beigel, Furst, and Rudich, who showed that if g is a $\mathsf{MAJ} \circ \mathsf{AC}^0_d$ circuit, then either g has size $\exp(n^{\Omega(1/d)})$, or else g disagrees with the parity function on a constant fraction of inputs [ABFR94]. Klivans combined this result with Yao's XOR Lemma to re-prove a strong (albeit not optimal) bound on the correlation between AC^0_d circuits and the parity function [Kli01]. Klivans' proof is the only prior work we are aware of that uses hardness amplification methods to prove an unconditional AC^0 circuit lower bound.

XOR lemmas for decision trees. Many prior works have studied XOR lemmas for various types of decision trees, along with the closely related "direct product" and "direct sum" problems [IRW94; BAN95; NRS99; Sha03; KŠW07; Špa08; AŠW09; JKS10; Dru12; She12; LR13; BK18; BB19; BKLS20]. However, as far as we are aware, we are the first to consider the case that we have hardness for all depths simultaneously.

1.6 Organization

After some preliminaries, we present our XOR lemma for decision trees (Lemma 1) in Section 3. Then, in Section 4, we present our correlation bound for depth reduction within AC^0 (Theorem 2), including our application to randomness extractors. Finally, in Section 5, we present our correlation bound for $\mathsf{MAJ}_n^{\oplus t}$ (Theorem 4).

2 Preliminaries

We write $\mathbb{N}$ to denote the set of non-negative integers.

2.1 Boolean Functions

In the introduction, we worked with functions $f \colon \{0,1\}^n \to \{0,1\}$. Going forward, it will be more convenient to encode a bit $b \in \{0,1\}$ as the value $(-1)^b$. Thus, we will work with functions $f \colon \{\pm 1\}^n \to \{\pm 1\}$. However, we will still use notation that is more typical for $\{0,1\}$-valued variables, namely:

$$\begin{aligned} \bigwedge_i x_i &:= \max_i x_i \\ \bigvee_i x_i &:= \min_i x_i \\ \bigoplus_i x_i &:= \prod_i x_i \\ \mathsf{MAJ}(x) &:= \operatorname{sign}\left(\sum_i x_i\right). \end{aligned}$$

We use the following notation for combining several copies of a Boolean function, generalizing the notation $h^{\oplus t}$ that we discussed in the introduction.

Definition 1 (Combining many copies of a Boolean function)**.** *Let $h \colon \{\pm 1\}^n \to \{\pm 1\}$ be a function, let $\square \in \{\oplus, \wedge, \vee\}$, and let $t \in \mathbb{N}$. We define $h^{\square t} \colon \{\pm 1\}^{nt} \to \{\pm 1\}$ by the rule*

$$h^{\square t}(x^{(1)}, \dots, x^{(t)}) = h(x^{(1)}) \mathbin{\square} \cdots \mathbin{\square} h(x^{(t)}).$$

We rely on the following upper bound on the size of AC^0 circuits computing the parity of a few bits.

Proposition 1 (AC^0_d upper bound for parity [Hås86b])**.** *Let $t \geq 1$ and $k \geq 2$ be integers. The parity function on t bits can be computed by an AC^0_k circuit of size $O(2^{t^{1/(k-1)}} \cdot t^{(k-2)/(k-1)})$. The output gate can be either an AND gate or an OR gate.*

We use the following notation to describe decision trees.

Definition 2 (Decision trees)**.** *For a function $f \colon \{\pm 1\}^n \to \{\pm 1\}$, we define* $\mathrm{DTDepth}(f)$ *to be the minimum depth of a decision tree computing f. In the other direction, for a parameter $D \in \mathbb{N}$, we define* $\mathrm{DTDepth}[D]$ *to be the class of all functions $f \colon \{\pm 1\}^n \to \{\pm 1\}$ that can be computed by depth-D decision trees. (The parameter n will always be clear from context.)*

2.2 Probability and Correlation

We denote random variables using boldface. We write $\mathbf{x} \sim \mu$ to indicate that the random variable $\mathbf{x}$ is sampled from the distribution μ. If $\mathbf{x}, \mathbf{x}'$ are discrete random variables taking values in Ω, then we consider the "total variation distance" between $\mathbf{x}$ and $\mathbf{x}'$ to be the maximum difference $|\Pr[\mathbf{x} \in S] - \Pr[\mathbf{x}' \in S]|$ over all $S \subseteq \Omega$. We use the following notation for product distributions.

Definition 3 (Tensor product of probability distributions)**.** *Let $\mu_1, \dots, \mu_t$ be probability distributions over the spaces $\Omega_1, \dots, \Omega_t$. Sample $\mathbf{x}_1 \sim \mu_1, \dots, \mathbf{x}_t \sim \mu_t$ independently. The* tensor product *$\mu_1 \otimes \cdots \otimes \mu_t$ is the probability distribution of $(\mathbf{x}_1, \dots, \mathbf{x}_t)$. As a special case, we define*

$$\mu^{\otimes t} = \underbrace{\mu \otimes \mu \otimes \cdots \otimes \mu}_{t \text{ copies}}.$$

We use the following standard definition to reason about average-case hardness of $\{\pm 1\}$-valued functions.

Definition 4 (Correlation)**.** *Let $g, h \colon \{\pm 1\}^n \to \{\pm 1\}$ be functions and let μ be a distribution over $\{\pm 1\}^n$. We define*

$$\mathsf{Corr}_\mu(g, h) = \mathop{\mathbb{E}}_{\mathbf{x} \sim \mu}[g(\mathbf{x}) \cdot h(\mathbf{x})].$$

More generally, if $\mathcal{C}$ is a class of functions $g \colon \{\pm 1\}^n \to \{\pm 1\}$, then we define

$$\mathsf{Corr}_\mu(\mathcal{C}, h) = \max_{g \in \mathcal{C}} \mathsf{Corr}_\mu(g, h).$$

If μ is omitted, then by default it is assumed to be the uniform distribution over $\{\pm 1\}^n$.

A bound $|\mathsf{Corr}(g, h)| \le \varepsilon$ is equivalent to the statement that g agrees with h on at most a $(1/2 + \varepsilon/2)$-fraction of inputs, because for any two $\{0, 1\}$-valued random variables $\mathbf{a}, \mathbf{b}$, we have $\Pr[\mathbf{a} = \mathbf{b}] = \frac{1}{2} + \frac{1}{2}\mathbb{E}[(-1)^{\mathbf{a}} \cdot (-1)^{\mathbf{b}}]$.

2.3 Generalized Restrictions

To formulate our XOR lemmas in the clearest possible way, we work with a notion of *generalized restrictions* that includes restrictions and projections as special cases. A generalized restriction, formally defined below, consists of an arbitrary "preprocessing" step that can be applied to a Boolean function of interest.

Definition 5 (Generalized restriction)**.** *A* generalized restriction *is a function $\pi \colon \{\pm 1\}^r \to \{\pm 1\}^n$. If $f \colon \{\pm 1\}^n \to \{\pm 1\}$ is a Boolean function, then we define $g|_\pi$ to be the composition $g \circ \pi$. That is, $g|_\pi \colon \{\pm 1\}^r \to \{\pm 1\}$ is given by $g|_\pi(x) = g(\pi(x))$.*

Traditional restrictions can be viewed as a special case of generalized restrictions as follows.

Definition 6 (Traditional restrictions as generalized restrictions)**.** *A* restriction *is a string $\rho \in \{+1, -1, \star\}^n$. For every $r \ge |\rho^{-1}(\star)|$, we identify ρ with a generalized restriction $\pi \colon \{\pm 1\}^r \to \{\pm 1\}^n$ as follows. Given $y \in \{\pm 1\}^r$, we let $\pi(y)$ be ρ, except that the i-th star is replaced with y_i for every i.*

Note that for convenience, we allow r (the number of variables that are "syntactically alive") to be greater than $|\rho^{-1}(\star)|$ (the number of variables that are "semantically alive"). Next, we consider *distributions* over generalized restrictions, and we explain how to interpret the tensor product of such distributions.

Definition 7 (Tensor product of generalized restriction distributions)**.** *Let $r, n \in \mathbb{N}$, and let $\mathcal{R}$ be a distribution over generalized restrictions $\boldsymbol{\pi} \colon \{\pm 1\}^r \to \{\pm 1\}^n$. Let $\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_t$ be independent samples from $\mathcal{R}$, and define $\vec{\boldsymbol{\pi}} \colon \{\pm 1\}^{rt} \to \{\pm 1\}^{nt}$ by concatenating, i.e.,*

$$\vec{\boldsymbol{\pi}}(y^{(1)}, \dots, y^{(t)}) = (\boldsymbol{\pi}_1(y^{(1)}), \dots, \boldsymbol{\pi}_t(y^{(t)})).$$

Then the tensor product $\mathcal{R}^{\otimes t}$ *is the distribution of the random variable $\vec{\boldsymbol{\pi}}$.*

2.4 Logarithmic Concavity

We recall the following standard definition.

Definition 8 (Log-concave)**.** *A function $f \colon [0, \infty) \to (0, \infty)$ is* log-concave *if $\log f$ is concave, i.e., for every $x, y \in \mathbb{R}_+$ and $\lambda \in (0, 1)$, we have $f(x)^\lambda \cdot f(y)^{1-\lambda} \leq f(\lambda x + (1 - \lambda)y)$.*

Note that if $f \colon [0, \infty) \to (0, \infty)$ is concave, then it is also log-concave, since $\log(x)$ is concave and monotone. Furthermore, by induction on t, if f is log-concave, then $\prod_{i=1}^t f(x_i) \leq f(\overline{x})^t$ where $\overline{x} = \frac{1}{t}\sum_{i=1}^t x_i$.

3 XOR Lemmas for Decision Trees

In this section, we present our XOR lemma for decision trees. We begin by stating a simple XOR lemma, in which the decision tree attempting to compute $h^{\oplus t}$ has the same depth as the decision tree attempting to compute h.

Lemma 2 (Basic XOR lemma for decision trees)**.** *Let $h_1, \dots, h_t \colon \{\pm 1\}^r \to \{\pm 1\}$ be functions, and define $h(y^{(1)}, \dots, y^{(t)}) = \prod_{i=1}^t h_i(y^{(i)})$. Let μ be a distribution over $\{\pm 1\}^r$. For every $D \in \mathbb{N}$, we have*

$$\mathsf{Corr}_{\mu^{\otimes t}}(h, \mathrm{DTDepth}[D]) \leq \prod_{i=1}^t \mathsf{Corr}_\mu(h_i, \mathrm{DTDepth}[D]).$$

We were unable to find a reference for the specific statement of Lemma 2, but it has no significant novelty. It is closely related to Shaltiel's analysis of "fair" decision trees [Sha03]. It can also be viewed as a special case of Claim 2 that we prove below. As discussed in subsection 1.4, Lemma 2 is sufficient for our analysis of depth-d approximators to AC^0_{d+k} circuits (Theorem 2). However, for our analysis of $\mathsf{MAJ}_n^{\oplus t}$ (Theorem 4), we need a more sophisticated XOR lemma, stated next.

Lemma 3 (XOR lemma for decision trees under robust hardness assumptions, general version)**.** *Let $h_1, \dots, h_t \colon \{\pm 1\}^r \to \{\pm 1\}$ be functions, and define $h(y^{(1)}, \dots, y^{(t)}) = \prod_{i=1}^t h_i(y^{(i)})$. Let $\mu_1, \dots, \mu_t$ be distributions over $\{\pm 1\}^r$, and define $\mu = \mu_1 \otimes \cdots \otimes \mu_t$. Let $\varepsilon \colon [0, \infty) \to (0, \infty)$ be a log-concave function, and assume that for every $i \in [t]$ and every $D \in \mathbb{N}$, we have*

$$\mathsf{Corr}_{\mu_i}(h_i, \mathrm{DTDepth}[D]) \leq \varepsilon(D).$$

Then for every $D \in \mathbb{N}$, we have

$$\mathsf{Corr}_\mu(h, \mathrm{DTDepth}[Dt/2]) \leq O(\varepsilon(D))^t.$$

The first step of the proof of Lemma 3 is the following claim, which enables us to relate the success probability of a tree to the success probabilities of its subtrees.

Claim 1 (Law of total correlation)**.** *Let $h, T, E\colon \{\pm1\}^r \to \{\pm1\}$. Let μ be a distribution over $\{0,1\}^r$ and let $j_* \in [r]$. For each $b \in \{\pm1\}$, let $p_b = \Pr_{\mathbf{y}\sim\mu}[E(\mathbf{y}) = b]$, and let μ^b be the conditional distribution $(\mathbf{y} \sim \mu \mid E(\mathbf{y}) = b)$. Suppose that T can be decomposed in the form*

$$T(y) = \begin{cases} T_{+1}(y) & \text{if } E(y) = +1 \\ T_{-1}(y) & \text{if } E(y) = -1 \end{cases}$$

for some $T_{+1}, T_{-1}\colon \{\pm1\}^r \to \{\pm1\}$. Then

$$\mathsf{Corr}_\mu(h, T) = \sum_{b\in\{\pm1\}} p_b \cdot \mathsf{Corr}_{\mu^b}(h, T_b).$$

Proof.

$$\begin{aligned} \mathsf{Corr}_\mu(h, T) &= \mathop{\mathbb{E}}_{\mathbf{y}\sim\mu}[h(\mathbf{y}) \cdot T(\mathbf{y})] \\ &= \sum_{b\in\{\pm1\}} p_b \cdot \mathop{\mathbb{E}}_{\mathbf{y}\sim\mu}[h(\mathbf{y}) \cdot T(\mathbf{y}) \mid E(\mathbf{y}) = b] \qquad \text{(Law of total expectation)} \\ &= \sum_{b\in\{\pm1\}} p_b \mathop{\mathbb{E}}_{\mathbf{y}\sim\mu^b}[h(\mathbf{y}) \cdot T_b(\mathbf{y})]. \end{aligned}$$

□

Next, we consider the following notion of "fair" decision trees due to Shaltiel [Sha03].

Definition 9 ($(D_1, \dots, D_t)$-fair decision trees [Sha03])**.** *Let $T\colon \{\pm1\}^{rt} \to \{\pm1\}$ be a decision tree and let $D_1, \dots, D_t \in \mathbb{N}$. We say that T is* $(D_1, \dots, D_t)$-fair *if for every input $\vec{y} = (y^{(1)}, \dots, y^{(t)}) \in (\{\pm1\}^r)^t$, for every $i \in [t]$, the computation $T(\vec{y})$ makes at most D_i queries to $y^{(i)}$.*

The key to proving Lemma 3 is to generalize Definition 9 to the case of a *set* of tuples $(D_1, \dots, D_t)$.

Definition 10 (Q-fair decision trees)**.** *Let $T\colon \{\pm1\}^{rt} \to \{\pm1\}$ be a decision tree and let $Q \subseteq \mathbb{N}^t$. We say that T is* Q-fair *if for every input $\vec{y} = (y^{(1)}, \dots, y^{(t)}) \in (\{\pm1\}^r)^t$, there is some tuple $(D_1, \dots, D_t) \in Q$ such that for every $i \in [t]$, the computation $T(\vec{y})$ makes at most D_i queries to $y^{(i)}$.*

We emphasize that the tuple $(D_1, \dots, D_t)$ is permitted to vary from one input $\vec{y}$ to another. Therefore, the fact that a tree is Q-fair does not necessarily imply that there is some $(D_1, \dots, D_t) \in Q$ such that the tree is $(D_1, \dots, D_t)$-fair. Given the concept of Q-fairness, it is relatively straightforward to prove the following claim by induction on the depth of T. The claim generalizes the analysis by Shaltiel [Sha03], who considered the case of $(D_1, \dots, D_t)$-fair decision trees and focused on the uniform distribution.

Claim 2 (XOR lemma for Q-fair decision trees with few query profiles)**.** *Let $h_1, \dots, h_t\colon \{\pm1\}^r \to \{\pm1\}$ be functions, and define $h(y^{(1)}, \dots, y^{(t)}) = \prod_{i=1}^t h_i(y^{(i)})$. Let $\mu_1, \dots, \mu_t$ be distributions over $\{\pm1\}^r$, and define $\mu = \mu_1 \otimes \cdots \otimes \mu_t$. Let $Q \subseteq \mathbb{N}^t$ and let $T\colon \{\pm1\}^{rt} \to \{\pm1\}$ be a Q-fair decision tree. Then*

$$\mathsf{Corr}_\mu(h, T) \le \sum_{(D_1,\dots,D_t)\in Q} \prod_{i=1}^t \mathsf{Corr}_{\mu_i}(h_i, \mathrm{DTDepth}[D_i]).$$

Proof. Assume without loss of generality that T never queries the same variable twice. For the base case, if T has depth 0, then T is a constant function, so

$$|\mathsf{Corr}_\mu(h, T)| = \prod_{i=1}^t \left| \mathop{\mathbb{E}}_{\mathbf{y}^{(i)}\sim\mu_i}[h_i(y^{(i)})] \right| = \prod_{i=1}^t \mathsf{Corr}_{\mu_i}(h_i, \mathrm{DTDepth}[0]).$$

Since T is Q-fair, Q must be nonempty. The lemma follows because $\mathsf{Corr}_{\mu_i}(h_i, \mathrm{DTDepth}[0]) \le \mathsf{Corr}_{\mu_i}(h_i, \mathrm{DTDepth}[D_i])$ for every $D_i \in \mathbb{N}$. For the inductive step, let $y_{j_*}^{(i_*)}$ be the variable queried by the root of the tree. Let T_{+1} and T_{-1} be the children of the root, corresponding to the cases $y_{j_*}^{(i_*)} = +1$ and $y_{j_*}^{(i_*)} = -1$ respectively. Define

$$Q' = \{(D_1, \dots, D_{i_*-1}, D_{i_*} - 1, D_{i_*+1}, \dots, D_t) : (D_1, \dots, D_t) \in Q \text{ and } D_{i_*} \neq 0\}.$$

Then T_{+1} and T_{-1} are both Q'-fair.

For each $b \in \{\pm 1\}$, define

$$p_b = \Pr_{\mathbf{y}^{(i_*)} \sim \mu_{i_*}} \left[\mathbf{y}_{j_*}^{(i_*)} = b \right].$$

Let $\mu_{i_*}^b$ be the conditional distribution $(\mathbf{y}^{(i_*)} \sim \mu_{i_*} \mid \mathbf{y}_{j_*}^{(i_*)} = b)$, and for $i \neq i_*$, let $\mu_i^b = \mu_i$. Let $\mu^b = \mu_1^b \otimes \cdots \otimes \mu_t^b$. By Claim 1 and the induction hypothesis, we have

$$\begin{aligned}
\mathsf{Corr}_\mu(h, T) &= \sum_{b \in \{\pm 1\}} p_b \cdot \mathsf{Corr}_{\mu^b}(h, T_b) \\
&\le \sum_{b \in \{\pm 1\}} p_b \cdot \sum_{(D_1, \dots, D_t) \in Q'} \prod_{i=1}^t \mathsf{Corr}_{\mu_i^b}(h_i, \mathrm{DTDepth}[D_i]) \\
&= \sum_{(D_1, \dots, D_t) \in Q'} \left(\sum_{b \in \{\pm 1\}} p_b \cdot \mathsf{Corr}_{\mu_{i_*}^b}(h_{i_*}, \mathrm{DTDepth}[D_{i_*}]) \right) \cdot \prod_{i \in [t], i \neq i_*} \mathsf{Corr}_{\mu_i}(h_i, \mathrm{DTDepth}[D_i]).
\end{aligned}$$

Now we bound the inner sum. By Claim 1, for any D_{i_*}, we have

$$\mathsf{Corr}_{\mu_{i_*}}(h_{i_*}, \mathrm{DTDepth}[D_{i_*} + 1]) \ge \sum_{b \in \{\pm 1\}} p_b \cdot \mathsf{Corr}_{\mu_{i_*}^b}(h_{i_*}, \mathrm{DTDepth}[D_{i_*}]),$$

because we can approximate h_{i_*} with respect to μ_{i_*} by first querying $y_{j_*}^{(i_*)}$ and then using optimal subtrees of depth D_{i_*}. For every $(D_1, \dots, D_t) \in Q'$, we have $(D_1, \dots, D_{i_*-1}, D_{i_*}+1, D_{i_*+1}, \dots, D_t) \in Q$. Therefore,

$$\mathsf{Corr}_\mu(h, T) \le \sum_{(D_1, \dots, D_t) \in Q} \prod_{i=1}^t \mathsf{Corr}_{\mu_i}(h_i, \mathrm{DTDepth}[D_i]). \qquad \square$$

Given Claim 2, our XOR lemma for decision trees under a robust hardness assumption (Lemma 3) readily follows, as we now show.

Proof of Lemma 3. Let $T \colon \{\pm 1\}^{rt} \to \{\pm 1\}$ be a decision tree of depth at most $Dt/2$. Let Q be the set of t-tuples $(D_1, \dots, D_t) \in \mathbb{N}^t$ such that (1) $D_1 + \cdots + D_t \le Dt$ and (2) D_i is an integer multiple of $\lceil D/2 \rceil$ for every i. We claim that T is Q-fair. Indeed, let $\vec{y} = (y^{(1)}, \dots, y^{(t)})$ be any input, and let D_i be the number of queries that $T(\vec{y})$ makes to $y^{(i)}$. Let D_i' be the smallest integer multiple of $\lceil D/2 \rceil$ such that $D_i \le D_i'$. Then $D_i' \le D_i + (\lceil D/2 \rceil - 1)$, and hence $D_1' + \cdots + D_t' \le Dt/2 + t \cdot (\lceil D/2 \rceil - 1) \le Dt$, showing that $(D_1', \dots, D_t') \in Q$.

Therefore, by Claim 2,

$$\mathsf{Corr}_\mu(h, T) \le \sum_{(D_1, \dots, D_t) \in Q} \prod_{i=1}^t \mathsf{Corr}_{\mu_i}(h_i, \mathrm{DTDepth}[D_i]).$$

For any $(D_1, \dots, D_t) \in Q$, we can define $(D'_1, \dots, D'_t)$ such that $D'_i \geq D_i$ and $D'_1 + \dots + D'_t$ is *exactly* Dt rather than being at most Dt. Then $\mathsf{Corr}_{\mu_i}(h_i, \mathrm{DTDepth}[D_i]) \leq \mathsf{Corr}_{\mu_i}(h_i, \mathrm{DTDepth}[D'_i])$, so

$$\begin{aligned}
\mathsf{Corr}_{\mu}(h, T) &\leq \sum_{(D_1,\dots,D_t)\in Q} \prod_{i=1}^{t} \mathsf{Corr}_{\mu_i}(h_i, \mathrm{DTDepth}[D'_i]) \\
&\leq \sum_{(D_1,\dots,D_t)\in Q} \prod_{i=1}^{t} \varepsilon(D'_i) \\
&\leq \sum_{(D_1,\dots,D_t)\in Q} \varepsilon(D)^t && \text{(Log-concavity)} \\
&= |Q| \cdot \varepsilon(D)^t.
\end{aligned}$$

To bound $|Q|$, observe that if $(D_1, \dots, D_t) \in Q$, then we can write $D_i = c_i \cdot \lceil D/2 \rceil$ for some nonnegative integers $c_1, \dots, c_t$. Furthermore, $Dt \geq \sum_i c_i \cdot \lceil D/2 \rceil \geq (D/2) \cdot \sum_i c_i$, so $c_1 + \dots + c_t \leq 2t$. Therefore, $|Q|$ is at most the number of ways that $2t$ can be partitioned into $t+1$ nonnegative integers, which is precisely $\binom{3t}{t}$. Thus,

$$\mathsf{Corr}_{\mu}(h, T) \leq \binom{3t}{t} \cdot \varepsilon(D)^t \leq O(\varepsilon(D))^t. \qquad \square$$

4 Correlation Bounds for Depth Reduction Within AC^0

In this section, we prove our result about the average-case hardness of AC^0_{d+k} circuits for AC^0_d circuits (Theorem 2). We begin by proving the basic version of our XOR lemma for the random simplification method. Then we review the basic structure of Håstad, Rossman, Servedio, and Tan's proof of the average-case depth hierarchy theorem [HRST17]. Finally, we combine the two to complete the proof of Theorem 2.

4.1 Basic XOR Lemma for the Random Simplification Method

Lemma 4 (XOR lemma for the random simplification method, basic version)**.** *Let $n, t, r, D \in \mathbb{N}$ and $\varepsilon, \delta > 0$. Let $h \colon \{\pm 1\}^n \to \{\pm 1\}$ and $g \colon \{\pm 1\}^{nt} \to \{\pm 1\}$ be Boolean functions, let $\mathcal{R}$ be a distribution over generalized restrictions $\boldsymbol{\pi} \colon \{\pm 1\}^r \to \{\pm 1\}^n$, let μ be a distribution over $\{\pm 1\}^r$, and assume the following.*

1. *(The distribution μ completes $\mathcal{R}$ to the uniform distribution) If we sample $\boldsymbol{\pi} \sim \mathcal{R}$ and $\mathbf{y} \sim \mu$ independently, then $\boldsymbol{\pi}(\mathbf{y})$ is a uniform random element of $\{\pm 1\}^n$.*

2. *(The function g simplifies under $\mathcal{R}^{\otimes t}$) We have*

$$\Pr_{\vec{\boldsymbol{\pi}} \sim \mathcal{R}^{\otimes t}} \left[\mathrm{DTDepth}(g|_{\vec{\boldsymbol{\pi}}}) > D \right] \leq \delta.$$

3. *(The function h retains structure under $\mathcal{R}$) We have*

$$\mathop{\mathbb{E}}_{\boldsymbol{\pi} \sim \mathcal{R}} \left[\mathsf{Corr}_{\mu}(h|_{\boldsymbol{\pi}}, \mathrm{DTDepth}[D]) \right] \leq \varepsilon.$$

Then $\mathsf{Corr}(g, h^{\oplus t}) \leq \varepsilon^t + \delta$.

Proof. Sample $\vec{\boldsymbol{\pi}} = (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_t) \sim \mathcal{R}^{\otimes t}$ and $\vec{\mathbf{y}} \sim \mu^{\otimes t}$ independently. Let $\mathbf{T}$ be $g|_{\vec{\boldsymbol{\pi}}}$ if $\mathrm{DTDepth}(g|_{\vec{\boldsymbol{\pi}}}) \leq D$; otherwise, let $\mathbf{T}$ be the constant-one function. Either way, $\mathbf{T}$ is a decision tree of depth at most D. Assumption 1 implies that $\vec{\boldsymbol{\pi}}(\vec{\mathbf{y}})$ is distributed uniformly over $\{\pm 1\}^{nt}$. Therefore,

$$\begin{aligned}
\mathsf{Corr}(h^{\oplus t}, g) &= \mathop{\mathbb{E}}_{\vec{\boldsymbol{\pi}}}\left[\mathsf{Corr}_{\mu^{\oplus t}}(h^{\oplus t}|_{\vec{\boldsymbol{\pi}}}, g|_{\vec{\boldsymbol{\pi}}})\right] && \text{(Assumption 1)}\\
&\leq \delta + \mathop{\mathbb{E}}_{\vec{\boldsymbol{\pi}}}\left[\mathsf{Corr}_{\mu^{\oplus t}}(h^{\oplus t}|_{\vec{\boldsymbol{\pi}}}, \mathbf{T})\right] && \text{(Assumption 2)}\\
&\leq \delta + \mathop{\mathbb{E}}_{\vec{\boldsymbol{\pi}}}\left[\prod_{i=1}^{t} \mathsf{Corr}_{\mu}(h|_{\boldsymbol{\pi}_i}, \mathrm{DTDepth}[D])\right] && \text{(Lemma 2)}\\
&= \delta + \left(\mathop{\mathbb{E}}_{\boldsymbol{\pi}\sim\mathcal{R}}\left[\mathsf{Corr}_{\mu}(h|_{\boldsymbol{\pi}}, \mathrm{DTDepth}[D])\right]\right)^t && \text{(Independence)}\\
&\leq \delta + \varepsilon^t && \text{(Assumption 3.)}
\end{aligned}$$

□

Remark 1 (Amplifying ε but not δ)**.** *When $t = 1$, the assumptions of Lemma 4 imply $\mathsf{Corr}(g, h) \leq \varepsilon + \delta$. When $t > 1$, we would therefore like to be able to say that $\mathsf{Corr}(g, h^{\oplus t})$ is at most roughly $(\varepsilon + \delta)^t$. Unfortunately, the conclusion of Lemma 4 has the weaker bound $\varepsilon^t + \delta$. To address this weakness, later we will prove a more sophisticated version of our XOR lemma (Lemma 8) with a stronger bound, albeit under stronger assumptions. As we will see, Lemma 4 is already sufficient for proving Theorem 2, because the bottleneck in Håstad, Rossman, Servedio, and Tan's correlation bound [HRST17] is in their "retaining structure" step rather than their "simplification" step. That is, in their argument, δ is much smaller than ε.*

Remark 2 (Simplification under $\mathcal{R}^{\otimes t}$ rather than $\mathcal{R}$)**.** *In Lemma 4, we assume that g simplifies under $\mathcal{R}^{\otimes t}$. In general, proving that circuits simplify under $\mathcal{R}^{\otimes t}$ could potentially be more difficult than proving that circuits simplify under $\mathcal{R}$. Thankfully, in the cases we consider, the distinction between $\mathcal{R}$ and $\mathcal{R}^{\otimes t}$ does not cause any serious difficulties.*

4.2 Review of Håstad, Rossman, Servedio, and Tan's Argument [HRST17]

Having proven our XOR lemma, to prove our correlation bound, our remaining job is to explain how Håstad, Rossman, Servedio, and Tan's argument [HRST17] fits into the assumptions of our XOR lemma. Let us therefore review their argument.

4.2.1 The Sipser Functions

To prove their average-case depth hierarchy theorem [HRST17], Håstad, Rossman, Servedio, and Tan used a hard function called the "Sipser function," which is a variant of the hard function used to prove the earlier worst-case hierarchy theorems [Sip83; Yao85; Hås86a]. The construction is parameterized by the depth of the hard function, denoted $d + 1$, and a parameter $m \in \mathbb{N}$ that determines the number of variables. To clarify the aspects of their argument that are important for us, we use slightly non-traditional notation to describe this construction below.

Definition 11 (The $\mathsf{USipser}$ functions [HRST17])**.** *For $m, d \geq 1$, we inductively define*

$$\mathsf{USipser}_{d,m} \colon \{\pm 1\}^{n_{d,m}} \to \{\pm 1\}$$

as follows. Let $f_{d,m}$ be a parameter that we will specify momentarily.

- *If $d = 1$, then $\mathsf{USipser}_{d,m}$ is an AND of $f_{d,m}$ distinct variables.*

- *If $d > 1$ and d is even, then* $\mathsf{USipser}_{d,m} = \mathsf{USipser}_{d-1,m}^{\vee f_{d,m}}$.[6]
- *If $d > 1$ and d is odd, then* $\mathsf{USipser}_{d,m} = \mathsf{USipser}_{d-1,m}^{\wedge f_{d,m}}$.

In each case, the fan-in parameter $f_{d,m}$ is chosen to be the smallest positive integer such that

$$\Pr_{\mathbf{x}}[\mathsf{USipser}_{d,m}(\mathbf{x}) = (-1)^d] \leq 2^{-2m}.$$

Thus, $\mathsf{USipser}_{d,m}$ *is a monotone read-once formula of depth d with AND gates adjacent to the input variables.*

Observe that under a uniform random input, $\mathsf{USipser}_{d,m}$ has acceptance probability roughly equal to either 2^{-2m} or $1 - 2^{-2m}$, depending on whether d is odd or even. The $\mathsf{BSipser}$ functions, defined below, correct this imbalance by adjusting only the fan-in of the output gate.[7]

Definition 12 (The $\mathsf{BSipser}$ functions [HRST17])**.** *For $m, d \geq 1$, we define*

$$\mathsf{BSipser}_{d+1,m} \colon \{\pm 1\}^{n'_{d+1,m}} \to \{\pm 1\}$$

as follows. Let $f'_{d+1,m}$ be a parameter that we will specify momentarily.

- *If $d+1$ is even, then* $\mathsf{BSipser}_{d+1,m} = \mathsf{USipser}_{d,m}^{\vee f'_{d+1,m}}$.
- *If $d+1$ is odd, then* $\mathsf{BSipser}_{d+1,m} = \mathsf{USipser}_{d,m}^{\wedge f'_{d+1,m}}$.

In each case, the fan-in parameter $f'_{d+1,m}$ is chosen to be the smallest positive integer such that

$$\Pr_{\mathbf{x}}[\mathsf{BSipser}_{d+1,m}(\mathbf{x}) = (-1)^{d+1}] \leq \frac{1}{2}.$$

Thus, $\mathsf{BSipser}_{d+1,m}$ *is a monotone read-once formula of depth $d+1$ with AND gates adjacent to the input variables.*

The hard function h in Håstad, Rossman, Servedio, and Tan's average-case depth hierarchy theorem (Theorem 1) is $\mathsf{BSipser}_{d+1,m}$ for a suitable parameter $m \approx \frac{\log n}{2d}$.

4.2.2 Håstad, Rossman, Servedio, and Tan's Random Projections [HRST17]

Håstad, Rossman, Servedio, and Tan prove the average-case hardness of $\mathsf{BSipser}$ using a carefully-engineered distribution over random projections. These projections are based on a special type of projection that we will call *fully-merging projections*. By definition, a fully-merging projection first applies a restriction and then merges all living variables to a single remaining variable. We give the definition below in terms of our "generalized restriction" formalism.

Definition 13 (Fully-merging projection)**.** *A* fully-merging projection *is a generalized restriction $\pi \colon \{\pm 1\} \to \{\pm 1\}^n$ such that for every $i \in [n]$, either $\pi(+1)_i = \pi(-1)_i$ ("variable i has been assigned a value"), or else $\pi(+1)_i = +1$ and $\pi(-1)_i = -1$ ("variable i is alive").*

[6] Recall the notation $h^{\vee f}$ from Definition 1.

[7] The "U" and "B" in $\mathsf{USipser}$ and $\mathsf{BSipser}$ stand for "Unbalanced" and "Balanced" respectively.

Let $n_{d,m}$ be the number of input variables to $\mathsf{USipser}_{d,m}$. In Håstad, Rossman, Servedio, and Tan's work, for each d and m, they carefully design a probability distribution $\mathcal{R}_{d,m}$ over fully-merging projections $\boldsymbol{\pi}\colon \{\pm 1\} \to \{\pm 1\}^{n_{d,m}}$. The inductive definition of $\mathcal{R}_{d,m}$ is fairly complicated, so we will refrain from reviewing the precise details. Instead, we merely cite the properties of these distributions that are important for our analysis.

The first crucial property of these projections is that plugging a uniform random input into $\mathsf{USipser}_{d,m}$ is equivalent to first applying a random projection, and then assigning the one remaining variable a random bit with a suitable bias.

Proposition 2 (Random projections complete to uniform [HRST17, Lemmas 7.3 and 8.4])**.** *For every $d, m \in \mathbb{N}$, there exists a distribution $\mu_{d,m}$ over $\{\pm 1\}$ such that if we sample $\boldsymbol{\pi} \sim \mathcal{R}_{d,m}$ and $\mathbf{y} \sim \mu_{d,m}$ independently, then $\boldsymbol{\pi}(\mathbf{y})$ is distributed uniformly over $\{\pm 1\}^{n_{d,m}}$.*

The next crucial property of Håstad, Rossman, Servedio, and Tan's random projections is that AC^0_d circuits *simplify* under tensor products of these projections.

Theorem 5 (Simplification of AC^0 circuits under random projections [HRST17])**.** *Let $m, d, \ell \in \mathbb{N}$ and assume that m is sufficiently large. Let $g\colon \{\pm 1\}^{n_{d,m}\cdot \ell} \to \{\pm 1\}$ be an AC^0_{d+1} circuit of size S with bottom fan-in at most $m/4$. Then*

$$\Pr_{\vec{\boldsymbol{\pi}} \sim \mathcal{R}_{d,m}^{\otimes \ell}}\left[\mathrm{DTDepth}(g|_{\vec{\boldsymbol{\pi}}}) > 2^{m/2-4}\right] \leq S \cdot 2^{-2^{m/2-4}}.$$

Theorem 5 is not quite stated in the form above anywhere in Håstad, Rossman, Servedio, and Tan's work [HRST17], but it follows from their analysis; see the proof of their "Theorem 10.1" for details.

Remark 3 (The role of ℓ)**.** *In Håstad, Rossman, Servedio, and Tan's analysis, they first fix d and m, and then they define projection distributions $\mathcal{R}^1, \dots, \mathcal{R}^{d-1}$, each of which operates on the input variables of $\mathsf{BSipser}_{d,m}$. They prove switching lemmas [HRST17, Lemmas 9.2 and 9.5] which analyze the effects of $\mathcal{R}^1, \dots, \mathcal{R}^{d-1}$ on AC^0_2 circuits, and an inductive argument demonstrates the effect of $\mathcal{R}^{d-1}$ on AC^0_{d-1} circuits (or more generally AC^0_d circuits with bounded bottom fan-in).*

The relationship between their notation $\mathcal{R}^1, \dots, \mathcal{R}^{d-1}$ and our notation $\mathcal{R}_{d,m}$ is given by $\mathcal{R}^i = \mathcal{R}_{i,m}^{\otimes \ell}$, where $\ell = \ell(i, m, d)$ is the number of gates in $\mathsf{BSipser}_{d,m}$ at distance i from the inputs. Their analysis is in fact applicable to $\mathcal{R}_{i,m}^{\otimes \ell}$ for an arbitrary *parameter ℓ, as indicated in Theorem 5. One quick way to convince oneself of this fact, without needing to go through their proofs line-by-line, is to observe that $\lim_{d\to\infty} \ell(i, m, d) = \infty$. Therefore, for any i, m, ℓ, there exists a large enough d such that our claim about $\mathcal{R}_{i,m}^{\otimes \ell}$ (Theorem 5) follows from Håstad, Rossman, Servedio, and Tan's analysis of $\mathsf{BSipser}_{d,m}$ [HRST17]. Here we are relying on the fact that the projection distribution $\mathcal{R}_{i,m}$ does not depend in any way on the depth d of the "ambient" $\mathsf{BSipser}$ formula, and the conclusions of Håstad, Rossman, Servedio, and Tan's switching lemmas [HRST17, Lemmas 9.2 and 9.5] have no dependence on d.*

Recall that the formula defining $\mathsf{BSipser}_{d+1,m}$ has top fan-in $f'_{d+1,m}$, and hence the total number of variables is $f'_{d+1,m} \cdot n_{d,m}$. The final crucial property of Håstad, Rossman, Servedio, and Tan's random projections is that $\mathsf{BSipser}_{d+1,m}$ *maintains structure* under $\mathcal{R}_{d,m}^{\otimes f'_{d+1,m}}$. Specifically, with high probability, after applying the projection, the circuit is still mildly hard for shallow decision trees with respect to the relevant distribution:

Proposition 3 (Sipser function maintains structure under random projections [HRST17])**.** *Let $d, m \in \mathbb{N}$, sample $\vec{\boldsymbol{\pi}} \sim \mathcal{R}_{d,m}^{\otimes f'_{d+1,m}}$, and let $\mu = \mu_{d,m}^{\otimes f'_{d+1,m}}$, where $\mu_{d,m}$ is the distribution from Proposition 2. Then*

$$\Pr_{\vec{\boldsymbol{\pi}}}\left[\mathsf{Corr}_{\mu}\left((\mathsf{BSipser}_{d+1,m})|_{\vec{\boldsymbol{\pi}}}, \mathrm{DTDepth}[2^{m/2}]\right) \leq O(2^{-m/4})\right] \geq 1 - O(2^{-m/2}).$$

Again, Proposition 3 does not appear in Håstad, Rossman, Servedio, and Tan's work [HRST17] in the form above, but it follows from their analysis; see the proof of their "Theorem 10.1."

4.3 Applying Our XOR Lemma

Plugging Håstad, Rossman, Servedio, and Tan's analysis into our XOR lemma yields the following correlation bound.

Theorem 6 (Correlation bound for parity of Sipser functions)**.** *Let $m, d, t, S \in \mathbb{N}$. Let $h = \mathsf{BSipser}_{d+1,m}$, and let n be the number of input variables to $h^{\oplus t}$. For every AC^0_d circuit $g\colon \{\pm 1\}^{nt} \to \{\pm 1\}$ of size S, we have*

$$\mathsf{Corr}(g, h^{\oplus t}) \leq O(2^{-m/4})^t + S \cdot 2^{-2^{m/2-4}}.$$

Proof. We apply Lemma 4 with $\mathcal{R} = \mathcal{R}_{d,m}^{\otimes f'_{d,m}}$ and $\mu = \mu_{d,m}^{\otimes f'_{d,m}}$. The first assumption of the lemma is satisfied by Proposition 2. The second assumption is satisfied with $D = 2^{m/2-4}$ and $\delta = S \cdot 2^{-2^{m/2-4}}$ by Theorem 5. The third assumption is satisfied with $\varepsilon = O(2^{-m/4})$ by Proposition 3. □

Finally, to prove our correlation bound for depth reduction within AC^0 (Theorem 2), essentially all that remains is to pick parameters.

Proof of Theorem 2. Define

$$m = \left\lfloor \frac{\log n}{3d} \right\rfloor \qquad \text{and } t = \left\lfloor \log^{k-2}\left(\frac{n}{\log^{k-3} n}\right) \right\rfloor.$$

Our hard function h is given by $h = \mathsf{BSipser}^{\oplus t}_{d+1,m}$. Note that due to our assumption on k, we have $\log^k n \leq n^{\alpha}$, and therefore $t \geq \Omega(\log n)^{k-2}$ and $t \leq n^{\alpha}$.

The function $\mathsf{BSipser}_{d+1,m}$ has $2^{2dm} \cdot m^d \cdot 2^{O(d)}$ variables [HRST17], which is bounded by $n^{2/3+O(\alpha)}$ by our choice of m and our assumption on d. Therefore, h has $n^{2/3+O(\alpha)}$ variables, which is at most n if we choose α to be a small enough constant (and we assume n is sufficiently large).

Recall that $\mathsf{BSipser}_{d+1,m}$ is a monotone read-once formula, so in particular it is an AC^0_{d+1} circuit of size $n^{2/3+O(\alpha)}$. The parity of t bits can be computed by an AC^0_{k-1} circuit of size $O(2^{t^{1/(k-2)}} \cdot t^{(k-3)/(k-2)})$ (Proposition 1), which is bounded by $O(n)$ by our choice of t. Therefore, h can be computed by an $\mathsf{AC}^0_{d+1+k-1}$ circuit of size $O(n) + 2t \cdot n^{2/3+O(\alpha)} = O(n)$.

Finally, let $g\colon \{\pm 1\}^n \to \{\pm 1\}$ be an AC^0_d circuit of size S. If $S \geq 2^{2^{m/2-5}}$, then we are done, because $2^{m/2-5} = n^{\Omega(1/d)}$. Assume now that $S \leq 2^{2^{m/2-5}}$. Then the last term of the correlation bound in Theorem 6 is at most $2^{-2^{m/2-5}} = 2^{-n^{\Omega(1/d)}}$, which is at most $2^{-\log^k n}$ by our assumption $dk \leq \frac{\alpha \log n}{\log\log n}$, provided we choose α to be a small enough constant. Meanwhile, the $O(2^{-m/4})^t$ term is clearly bounded by $2^{-\frac{1}{d}\cdot\Omega(\log n)^{k-1}}$, completing the proof. □

Remark 4 (Depth complexity of our hard function)**.** *In the proof above, we argued that the hard function h can be computed by a linear-size AC^0_{d+k} circuit. There is a temptation to try to argue that a circuit of depth $d+k-1$ suffices, by merging the bottom layer of the parity circuit with the top gates of the* $\mathsf{BSipser}$ *formulas. Unfortunately, this argument does not work. The issue is that parity is not monotone, so the parity circuit has negations. After propagating the negations down through the* $\mathsf{BSipser}$ *formulas, some of the* $\mathsf{BSipser}$ *formulas have AND gates on top whereas others have OR gates on top.*

4.4 Application: Extractors for AC^0-Recognizable Sources

In this section, to illustrate the qualitative difference between our inverse-quasipolynomial correlation bound (Theorem 2) and Håstad, Rossman, Servedio, and Tan's greater-than-$(1/n)$ correlation bound (Theorem 1), we explain how to use our new correlation bound to construct new *seedless randomness extractors*. Specifically, we construct extractors for *recognizable sources*, a concept introduced by Shaltiel [Sha11]. We review the definition below.

Definition 14 (Extractor for recognizable sources [Sha11])**.** *Let $n \in \mathbb{N}$ and let $\mathcal{C}$ be a class of Boolean functions $g\colon \{\pm 1\}^n \to \{\pm 1\}$. For each $g \in \mathcal{C}$, let $\mathbf{U}_g$ denote the uniform distribution over $g^{-1}(-1)$. Let $\kappa \in \mathbb{N}$, and let $\varepsilon > 0$. A (κ, ε)-*extractor for sources recognizable by $\mathcal{C}$ *is a function $E\colon \{\pm 1\}^n \to \{\pm 1\}^m$ such that for every $g \in \mathcal{C}$, if $\mathbf{U}_g$ has min-entropy*[8] *at least κ, then $E(\mathbf{U}_g)$ is ε-close to the uniform distribution over $\{\pm 1\}^m$ in total variation distance.*

Shaltiel's initial motivation for studying extractors for recognizable sources was an application to typically-correct derandomization [Sha11]. Later, Applebaum, Artemenko, Shaltiel, and Yang used extractors for recognizable sources to construct incompressible functions [AASY16]. We believe that the concept of an extractor for recognizable sources is interesting in its own right. We construct an extractor, computable by near-linear-size $\mathsf{AC}^0_{d+O(1)}$ circuits, for sources that are recognizable by AC^0_d circuits of size up to $\exp(n^{\Theta(1/d)})$.

Theorem 7 (Extractors for (AC^0_d)-recognizable sources, computable by AC^0_{d+k} circuits)**.** *For every constant $\gamma > 0$, there is a constant $\alpha > 0$ such that for every $n, d, k \in \mathbb{N}$ with $k \geq 4$ and $dk \leq \frac{\alpha \log n}{\log\log n}$, there exists an explicit AC^0_{d+k} circuit*[9] *$E\colon \{\pm 1\}^n \to \{\pm 1\}^{n-n^{\gamma}}$ of size $n^{1+\gamma}$ such that E is an $(n-\Delta, 2^{-\Delta})$-extractor for sources that are recognizable by AC^0_d circuits of size at most S, where $\Delta = \frac{1}{d} \cdot \Omega(\log n)^{k-2}$ and $S = \exp(n^{\Omega(1/d)})$.*

Remark 5 (Prior extractors for AC^0-recognizable sources)**.** *Li and Zuckerman constructed extractors for sources recognizable by AC^0_d circuits based on the hardness of the parity function [LZ19], improving a prior construction by Shaltiel [Sha11]. The entropy and error parameters of their extractor are superior to ours. However, if we wish to extract from sources recognized by circuits of size $S = \exp(n^{\Omega(1/d)})$, then computing their extractor involves computing the parity of $n^{\Omega(1)}$ bits, and hence their extractor has much higher computational complexity than ours.*[10]

Another approach for constructing an extractor for (AC^0_d)-recognizable sources would be to combine Håstad, Rossman, Servedio, and Tan's weak correlation bound (Theorem 1) with Shaltiel's

[8] By definition, a random variable $\mathbf{x}$ has min-entropy at least κ if $\Pr[\mathbf{x} = x] \leq 2^{-\kappa}$ for every x.

[9] Note that E is a multi-output function. An m-output AC^0_d circuit is a list of m single-output AC^0_d circuits. The size of the m-output circuit is the sum of the sizes of the constituent single-output circuits.

[10] Li and Zuckerman [LZ19] and Shaltiel [Sha11] also consider sources recognizable by smaller AC^0 circuits. In this setting, their extractors have lower computational complexity. For example, if we are only trying to extract from sources recognized by circuits of size $S = \mathrm{poly}(n)$, then their extractors can indeed be implemented by polynomial-size AC^0 circuits, because such circuits can compute the parity of $\operatorname{polylog} n$ bits.

non-PRG-based extractor construction [Sha11]. The main weakness of this approach is that the extractor's output length would only be $n^{O(1/d)}$, even though the extractor would require an input with $n - \Theta(\frac{1}{d}\log n)$ bits of entropy. In contrast, note that our extractor has an additive *entropy loss of only n^{β} bits.*

The first step of the proof of Theorem 7 is to use the Nisan-Wigderson framework [NW94] to convert our correlation bound (Theorem 2) into a PRG. We emphasize that this step is possible only because our correlation bound is less than $1/n$. We begin by recalling the formal definition of a PRG.

Definition 15 (PRGs)**.** *Let $n \in \mathbb{N}$, let $\mathcal{C}$ be a class of functions $g\colon \{\pm 1\}^n \to \{\pm 1\}$, let $G\colon \{\pm 1\}^s \to \{\pm 1\}^n$ be a function, and let $\varepsilon > 0$. We say that G is a* PRG that fools $\mathcal{C}$ with error ε *if for every $g \in \mathcal{C}$, we have*

$$\left| \mathop{\mathbb{E}}_{\mathbf{x}\in\{\pm 1\}^n}[g(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x}\in\{\pm 1\}^s}[g(G(\mathbf{x}))] \right| \leq \varepsilon.$$

The parameter s is called the seed length *of G.*

We are interested in *seed-extending* PRGs, a concept introduced by Kinne, van Melkebeek, and Shaltiel [KMS12].

Definition 16 (Seed-extending PRGs [KMS12])**.** *A PRG $G\colon \{\pm 1\}^s \to \{\pm 1\}^n$ is* seed-extending *if there is some function $G'\colon \{\pm 1\}^s \to \{\pm 1\}^{n-s}$ such that for every seed x, we have*

$$G(x) = (x, G'(x)).$$

By applying the standard Nisan-Wigderson framework to our correlation bound, we get a PRG with the following parameters.

Theorem 8 (Seed-extending PRG fooling AC^0_d, computable in AC^0_{d+k})**.** *For every $n, d, k, S \in \mathbb{N}$ with $k \geq 4$, $S \geq n$ and for every $\varepsilon > 0$, there exists a seed-extending PRG $G\colon \{\pm 1\}^s \to \{\pm 1\}^n$ that fools AC^0_d circuits of size S with error ε and seed length*

$$s = (\log S)^{O(d)} + \exp\left(O\left((d \cdot \log(n/\varepsilon))^{1/(k-2)}\right)\right) + \exp\left(O(dk \cdot \log(dk))\right),$$

such that for every $i \in [n]$, there is an explicit AC^0_{d+k} circuit $G_i\colon \{\pm 1\}^s \to \{\pm 1\}$ of size s computing the function $G_i(x) = G(x)_i$.

Proof. Let $h\colon \{\pm 1\}^r \to \{\pm 1\}$ be our AC^0_{d+k} circuit such that $\mathsf{Corr}(g, h) \leq \delta$ for every AC^0_{d+1} circuit g of size S_0, where $\delta = \exp(-\Omega(\frac{1}{d}\log^{k-2} r))$, $S_0 = 2^{r^{\Omega(1/d)}}$, and the parameter r will be specified later. Let $I_1, \ldots, I_n \subseteq [r^2]$ be a polynomial-time-constructible family of sets such that $|I_i| = r$ for every i and $|I_i \cap I_j| < \log n$ whenever $i \neq j$; such families exist provided r is a power of two [Nis91; NW94]. Let $s = r^2$, and define $G\colon \{\pm 1\}^s \to \{\pm 1\}^{s+n}$ by

$$G(x) = (x, h(x|_{I_1}), h(x|_{I_2}), \ldots, h(x|_{I_{n-s}})).$$

The standard Nisan-Wigderson analysis [NW94] shows that G fools AC^0_d circuits of size $S_0 - 2n^2$ with error δn. We must choose r large enough to satisfy three constraints:

- We must ensure that $(d+1)\cdot(k-1) \leq \frac{\alpha \log r}{\log\log r}$ as required by Theorem 2.
- We must ensure that $S_0 - 2n^2 \geq S$ so that G fools all AC^0_d circuits of size S.

- We must ensure that $\delta n \leq \varepsilon$ so that G has error at most ε.

We can satisfy all three of those conditions with a suitable choice

$$r = \exp(O(d \cdot k \cdot \log(dk))) + (\log S)^{O(d)} + \exp\left(O(d \log(n/\varepsilon))^{1/(k-2)}\right).$$

Clearly, each individual output bit of G can be computed by an explicit AC^0_{d+k} circuit of size $O(r) \leq s$. □

Remark 6 (Strongly explicit PRGs)**.** *In Theorem 8, we construct a separate circuit for each index $i \in [n]$. With slightly more effort and slightly worse parameters, one can construct a single circuit of size s computing the function $\overline{G}(x, i) := G(x)_i$.*

Remark 7 (Prior PRGs in AC^0 that fool AC^0 circuits)**.** *Our PRG should be compared to several prior unconditional PRGs. If we focus on fooling* polynomial-size *AC^0_d circuits, then classic PRGs such as Nisan's PRG [Nis91] can be computed by polynomial-size $\mathsf{AC}^0_{d+O(1)}$ circuits, and in fact Viola constructed a PRG that is computable by polynomial-size AC^0_2 circuits [Vio12]. In this paper, we are primarily interested in the challenge of fooling much larger AC^0_d circuits, namely circuits of size $\exp(n^{\Omega(1/d)})$. In this regime, prior work by Mossel, Shpilka, and Trevisan is relevant [MST06]. They constructed "small-bias" PRGs that are computable in NC^0, i.e., each bit of the PRG's output depends on only a constant number of bits of the seed. Their PRGs have exponentially small bias, and every δ-biased distribution is $(\delta \cdot n^k)$-close to a k-wise independent distribution [AGM03], and k-wise independent distributions ε-fool AC^0_d circuits of size S when $k = (\log S)^{O(d)} \cdot \log(1/\varepsilon)$ [Baz09; Raz09; Bra10; Tal17; HS19]. As a result, the Mossel-Shpilka-Trevisan PRG [MST06] fools AC^0_d circuits of size $\exp(n^{\Omega(1/d)})$, similar to what we get using the Nisan-Wigderson framework [NW94]. However, a crucial distinction is that because we construct our PRG with the Nisan-Wigderson framework, our PRG is seed-extending. This enables us to construct our randomness extractor. The Mossel-Shpilka-Trevisan PRG [MST06] is not seed-extending.*

Next, we use the following reduction due to Li and Zuckerman [LZ19] (improving earlier work by Kinne, van Melkebeek, and Shaltiel [KMS12]) showing how to convert any seed-extending PRG into an extractor for recognizable sources.

Theorem 9 (Seed extending PRG $\implies$ extractor for recognizable sources [LZ19, Theorem 8])**.** *Let $n \in \mathbb{N}$ and let $\mathcal{C}$ be a class of functions $g\colon \{\pm 1\}^n \to \{\pm 1\}$. Assume that $\mathcal{C}$ is "flip-invariant," i.e., for every $g \in \mathcal{C}$ and every $y \in \{\pm 1\}^n$, the function $h(x) := g(x_1 y_1, \dots, x_n y_n)$ is also in $\mathcal{C}$. Let $G\colon \{\pm 1\}^s \to \{\pm 1\}^n$ be a seed-extending ε-PRG for $\mathcal{C}$, namely $G(x) = (x, G'(x))$, and define $E\colon \{\pm 1\}^n \to \{\pm 1\}^{n-s}$ by the formula $E(x, y) = G'(x) \oplus y$ (where $\oplus$ denotes bitwise product of $\{\pm 1\}$ values). Then for every $\Delta > 0$, the function E is an $(n - \Delta, 2^{\Delta}\varepsilon)$-extractor for $\mathcal{C}$-recognizable sources.*

Theorem 7 follows by combining our PRG (Theorem 8) with Li and Zuckerman's reduction (Theorem 9):

Proof of Theorem 7. Let $G\colon \{\pm 1\}^s \to \{\pm 1\}^n$ be the seed-extending $(2^{-2\Delta})$-PRG for AC^0_d circuits of size S from Theorem 8. Choose $S = 2^{n^{\Omega(1/d)}}$ and $\Delta = \frac{1}{d} \cdot \Omega(\log n)^{k-2} = \omega(\log n)$ such that the seed length s is at most n^{β}. Since G is seed-extending, we can write it as $G(x) = (x, G'(x))$. The extractor is given by $E(x, y) = G'(x) \oplus y$. By Theorem 9, E is an $(n - \Delta, 2^{-\Delta})$-extractor for sources recognizable by AC^0_d circuits of size S. All that remains is to verify the computational complexity of E.

Each output bit of G' can be computed by an AC^0_{d+k} circuit of size n^{β}. The i-th output bit of $E(x, y)$ is given by $G'(x)_i \oplus y_i$. Naively, this looks like a circuit of depth $d + k + 2$. To avoid paying the "+2" penalty for the XOR with y_i, we recall the structure of the circuit computing G'. The top $k - 2$ layers of the circuit computing G' simply consist of a circuit computing the parity of $t = \Theta(\log^{k-3} n)$ bits. We can incorporate one more input variable (y_i) into this parity computation without increasing the depth of the circuit and with no asymptotic increase in the size. □

5 Correlation Bound for XOR of Majority

In this section, we prove our correlation bound for the $\mathsf{MAJ}_n^{\oplus t}$ function (Theorem 4). More generally, recall that a function $h \colon \{\pm 1\}^n \to \{\pm 1\}$ is *symmetric* if $h(x)$ depends only on the number of $+1$ values in x. We prove the following correlation bound for $h^{\oplus t}$ where h is any symmetric function:

Theorem 10 (XORing amplifies hardness for symmetric functions)**.** *Let $n, t, d, S \in \mathbb{N}$. Let $h \colon \{\pm 1\}^n \to \{\pm 1\}$ be a symmetric function, and let $g \colon \{\pm 1\}^{nt} \to \{\pm 1\}$ be an AC^0_d circuit of size S. Then*

$$\mathsf{Corr}(g, h^{\oplus t}) \leq \left(O\left(\left| \mathop{\mathbb{E}}_{\mathbf{x} \in \{\pm 1\}^n} [h(\mathbf{x})] \right| \right) + \frac{O(\log S)^{d-1} \cdot \sqrt{\log n}}{\sqrt{n}} \right)^t .$$

Theorem 10 implies Theorem 4, because if $h = \mathsf{MAJ}_n$, then $|\mathbb{E}_{\mathbf{x}}[h(\mathbf{x})]|$ is either 0 (if n is odd) or $O(1/\sqrt{n})$ (if n is even).

To prove Theorem 10, we begin by presenting a new random-restrictions-based proof that MAJ_n (and more generally any near-balanced symmetric function) is moderately hard for AC^0_d circuits. Then, we prove a more sophisticated variant of our XOR lemma for the random simplification method based on our more sophisticated XOR lemma for decision trees (Lemma 3). Combining these two ingredients will complete the proof.

5.1 Correlation Bound for Majority via a Random Simplification Argument

5.1.1 AC^0 circuits simplify under suitable random restrictions

We use the following notation for truly random restrictions.

Definition 17 (Truly random restriction $\mathcal{R}_{p,n}$)**.** *Let $\mathcal{R}_{p,n}$ denote the following distribution over $\{+1, -1, \star\}^n$. To sample $\boldsymbol{\rho} \sim \mathcal{R}_{p,n}$, for each coordinate $i \in [n]$ independently, we set*

$$\boldsymbol{\rho}_i = \begin{cases} \star & \text{with probability } p \\ +1 & \text{with probability } (1-p)/2 \\ -1 & \text{with probability } (1-p)/2. \end{cases}$$

To prove our correlation bound for the majority function, we use a slightly different distribution over restrictions. To define this distribution, let us introduce some convenient notation that we will use throughout this section. For a string $\rho \in \{+1, -1, \star\}^*$, we define $\Sigma(\rho) = |\rho^{-1}(+1)| - |\rho^{-1}(-1)|$. We will frequently apply this definition in the special case $x \in \{\pm 1\}^n$; in this case $\Sigma(x) = \sum_{i=1}^n x_i$. If $n \in \mathbb{N}$, $p > 0$, and $(1-p)n$ is an even integer, then we define

$$\begin{aligned} E_{p,n} &= \{x \in \{\pm 1\}^n : |\Sigma(x)| > pn\} \\ R_{p,n} &= \left\{x \in \{+1, -1, \star\}^n : |x^{-1}(\star)| = pn \text{ and } \Sigma(x) = 0\right\}. \end{aligned}$$

Definition 18 (Random restrictions tailored to majority)**.** *Let $n \in \mathbb{N}$ and $p > 0$ such that $(1-p)n$ is an even positive integer. We define $\widetilde{\mathcal{R}}_{p,n}$, a distribution over generalized restrictions $\boldsymbol{\pi}\colon \{\pm 1\}^{pn} \to \{\pm 1\}^n$, by the following sampling procedure.*

1. *With probability $|E_{p,n}|/2^n$, output a uniform random element of $E_{p,n}$ (an assignment to all n variables).*
2. *With probability $1 - |E_{p,n}|/2^n$, output a uniform random element of $R_{p,n}$ (a balanced assignment to $(1-p)n$ variables).*

Note that if we sample $\boldsymbol{\pi} \sim \widetilde{\mathcal{R}}_{p,n}$ and apply it to a function $f\colon \{\pm 1\}^n \to \{\pm 1\}$, then according to our definitions, $f|_{\boldsymbol{\pi}}$ is a function on $\{\pm 1\}^{pn}$, even in the rare case that $\boldsymbol{\pi}$ assigns values to all variables (in which case $f|_{\boldsymbol{\pi}}$ is constant). See Definition 6 for a reminder of how elements of $E_{p,n}$ and elements of $R_{p,n}$ can all be interpreted as generalized restrictions $\pi\colon \{\pm 1\}^{pn} \to \{\pm 1\}^n$.

To analyze $\widetilde{\mathcal{R}}_{p,n}$, we use the following estimate of the maximum value of the binomial distribution probability mass function, which follows from Stirling's approximation.

Proposition 4 (Probability of getting exactly pn heads among n tosses of a coin with bias p)**.** *Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be independent $\{0,1\}$-valued random variables such that $\Pr[\mathbf{x}_i = 1] = p$ for every i, where $p \in (0, 1/2]$ and pn is an integer. Then $\Pr[\sum_i \mathbf{x}_i = pn] = \Theta(1/\sqrt{pn})$.*

Now we show that AC^0_d circuits simplify under (tensor products of) this distribution $\widetilde{\mathcal{R}}_{p,n}$. This follows readily as a consequence of prior work studying $\mathcal{R}_{p,n}$:

Lemma 5 (Simplification of AC^0_d circuits under restrictions tailored to the majority function)**.** *Let n, t, S, d be positive integers, and let $g\colon \{\pm 1\}^{nt} \to \{\pm 1\}$ be an AC^0_d circuit. There exists a value $p = 1/O(\log S)^{d-1}$ such that $(1-p)n$ is an even positive integer and for every $D \in \mathbb{N}$,*

$$\Pr_{\vec{\boldsymbol{\pi}} \sim \widetilde{\mathcal{R}}_{p,n}^{\otimes t}}[\mathrm{DTDepth}(g|_{\vec{\boldsymbol{\pi}}}) > D] \le O(n)^t \cdot 2^{-D}.$$

Proof. First we prove the theorem for the case that $\vec{\boldsymbol{\pi}}$ is sampled uniformly from $R_{p,n}^t$. We have

$$\begin{aligned}\Pr_{\vec{\boldsymbol{\pi}} \in R_{p,n}^t}[\mathrm{DTDepth}(g|_{\vec{\boldsymbol{\pi}}}) > D] &= \Pr_{\vec{\boldsymbol{\rho}} \sim \mathcal{R}_{p,nt}}[\mathrm{DTDepth}(g|_{\vec{\boldsymbol{\rho}}}) > D \mid \vec{\boldsymbol{\rho}} \in R_{p,n}^t] \\ &\le \frac{\Pr_{\vec{\boldsymbol{\rho}} \sim \mathcal{R}_{p,nt}}[\mathrm{DTDepth}(g|_{\vec{\boldsymbol{\rho}}}) > D]}{\Pr_{\vec{\boldsymbol{\rho}} \sim \mathcal{R}_{p,nt}}[\vec{\boldsymbol{\rho}} \in R_{p,n}^t]}.\end{aligned}$$

First consider the numerator. Building on Håstad's switching lemma and multi-switching lemma [Hås86a; Hås14], Rossman showed that $\Pr_{\vec{\boldsymbol{\rho}} \sim \mathcal{R}_{q,nt}}[\mathrm{DTDepth}(g|_{\vec{\boldsymbol{\rho}}}) > D] \le (q \cdot O(\log S)^{d-1})^D$ [Ros17], which is at most 2^{-D} for a suitable choice $q = 1/O(\log S)^{d-1}$. In the extreme case $q \le 3/n$, we let $p = 1/n$ or $p = 2/n$, whichever ensures that $(1-p)n$ is an even positive integer. Then every restriction $\vec{\boldsymbol{\pi}} \in R_{p,n}^t$ leaves at most $2t$ variables alive, so trivially $\mathrm{DTDepth}(g|_{\boldsymbol{\rho}}) \le 2t$, completing the proof. In the more important case that $q > 3/n$, pick $p \in [q - 2/n, q]$ to ensure that $(1-p)n$ is even positive integer and note that $p = \Omega(q)$.

Now consider the denominator. The probability that a restriction drawn from $\mathcal{R}_{p,n}$ lands in $R_{p,n}$ is $\Theta(\frac{1}{n\sqrt{p}}) \ge \Omega(1/n)$ by Proposition 4. Therefore, $\Pr_{\vec{\boldsymbol{\rho}} \sim \mathcal{R}_{p,nt}}[\vec{\boldsymbol{\rho}} \in R_{p,n}^t] \ge \Omega(1/n)^t$, completing the proof for this case. (A similar argument was used by Oliveira, Santhanam, and Srinivasan [OSS19].)

Now we prove the theorem for the case that $\vec{\boldsymbol{\pi}}$ is sampled from $\widetilde{\mathcal{R}}_{p,n}^{\otimes t}$. We can sample from $\widetilde{\mathcal{R}}_{p,n}^{\otimes t}$ by the following procedure.

1. Sample $\mathbf{e}_1, \ldots, \mathbf{e}_t \in \{0,1\}$ independently, where $\Pr[\mathbf{e}_i = 1] = |E_{p,n}|/2^n$ for each i.
2. For every i such that $\mathbf{e}_i = 1$, pick a uniform random element of $E_{p,n}$ and assign it to the i-th block of the variables. Let t' be the number of i such that $\mathbf{e}_i = 0$.
3. Sample $\vec{\boldsymbol{\pi}}' \in R_{p,n}^{t'}$ uniformly at random and use it to assign values to $(1-p)n$ variables in each remaining block.

After performing steps 1 and 2, we have an AC_d^0 circuit of size S on $n \cdot t'$ variables, so applying our analysis of $R_{p,n}^{t'}$ completes the proof. □

5.1.2 Completion to the uniform distribution

Recall that $\widetilde{\mathcal{R}}_{p,n}$ leaves pn variables "syntactically alive," i.e., $\widetilde{\mathcal{R}}_{p,n}$ is a distribution over generalized restrictions $\boldsymbol{\pi}\colon \{\pm 1\}^{pn} \to \{\pm 1\}^n$. (Rarely, $\boldsymbol{\pi}$ is a constant function, in which case we could say that the pn variables are not "semantically alive.") Now we define a distribution $\mu_{p,n}$ over $\{\pm 1\}^{pn}$ that "completes" $\widetilde{\mathcal{R}}_{p,n}$ to the uniform distribution.

Definition 19 (Residual distribution for $\widetilde{\mathcal{R}}_{p,n}$)**.** *Let $n \in \mathbb{N}$ and $p > 0$ such that $(1-p)n$ is an even positive integer. We define $\mu_{p,n}$, a distribution over $\{\pm 1\}^{pn}$, by the following sampling procedure.*

1. *Sample $\mathbf{x} \in \{\pm 1\}^n \setminus E_{p,n}$ uniformly at random.*
2. *Sample $\mathbf{y}$ uniformly at random from the set $\{y \in \{\pm 1\}^{pn} : \Sigma(y) = \Sigma(\mathbf{x})\}$.*
3. *Output $\mathbf{y}$.*

Lemma 6 (Completion to the uniform distribution)**.** *Let $n \in \mathbb{N}$ and $p > 0$ such that $(1-p)n$ is an even positive integer. Sample $\boldsymbol{\pi} \sim \widetilde{\mathcal{R}}_{p,n}$ and $\mathbf{y} \sim \mu_p$ independently. Then $\boldsymbol{\rho}(\mathbf{y})$ is distributed uniformly over $\{\pm 1\}^n$.*

Proof. Looking at the definitions, we see that $\boldsymbol{\rho}(\mathbf{y})$ can be sampled by the following procedure.

1. Sample $\mathbf{x} \in \{\pm 1\}^n$ uniformly at random.
2. If $\mathbf{x} \in E_{p,n}$, then output $\mathbf{x}$. Otherwise:
3. Permute the coordinates of $\mathbf{x}$ to form a string $\mathbf{x}'$ consisting of $(1-p)n/2$ many $+1$ values, followed by $(1-p)n/2$ many -1 values, followed by the remaining ± 1 values in an arbitrary order.
4. Sample a uniform random permutation $\boldsymbol{\sigma}\colon [n] \to [n]$, and output the string $(\mathbf{x}'_{\boldsymbol{\sigma}(1)}, \ldots, \mathbf{x}'_{\boldsymbol{\sigma}(n)})$.

(The indices $\boldsymbol{\sigma}^{-1}(1), \ldots, \boldsymbol{\sigma}^{-1}((1-p)n)$ are the positions that are assigned values by $\boldsymbol{\pi}$.) This procedure clearly generates a uniform random string. □

5.1.3 Majority retains structure under restrictions

Now we show that with high probability over $\boldsymbol{\pi} \sim \widetilde{\mathcal{R}}_p$, the function $(\mathsf{MAJ}_n)|_{\boldsymbol{\pi}}$ is moderately hard for shallow decision trees with respect to $\mu_{p,n}$. More generally, we will bound $\mathsf{Corr}_{\mu_{p,n}}(h, \mathrm{DTDepth}[D])$ for any symmetric function h. We rely on the following bound, which can be proven using Pinsker's inequality.

Proposition 5 (Fair coin vs. biased coin)**.** *Sample* $\mathbf{x}_1, \dots, \mathbf{x}_D \in \{\pm 1\}$ *independently and uniformly at random. Let* $\varepsilon \in (0, 1/2]$*, and sample* $\widetilde{\mathbf{x}}_1, \dots, \widetilde{\mathbf{x}}_D \in \{\pm 1\}$ *independently such that* $\Pr[\widetilde{\mathbf{x}}_i = +1] = 1/2 + \varepsilon$ *for every* i*. Then the total variation distance between* $\mathbf{x} := (\mathbf{x}_1, \dots, \mathbf{x}_D)$ *and* $\widetilde{\mathbf{x}} := (\widetilde{\mathbf{x}}_1, \dots, \widetilde{\mathbf{x}}_D)$ *is at most* $O(\varepsilon\sqrt{D})$*.*

We also rely on the following bound by Diaconis and Freedman [DF80] relating sampling without replacement to sampling with replacement.

Theorem 11 (Sampling with vs. without replacement [DF80])**.** *Let* Ω *be a finite alphabet, let* $y \in \Omega^r$*, and let* $D \le r$*. Sample indices* $\mathbf{i}_1, \dots, \mathbf{i}_D \in [r]$ *uniformly at random without replacement, and sample indices* $\widetilde{\mathbf{i}}_1, \dots, \widetilde{\mathbf{i}}_1 \in [r]$ *uniformly at random with replacement (i.e., uniformly and independently). Then the total variation distance between* $\mathbf{x} := (y_{\mathbf{i}_1}, \dots, y_{\mathbf{i}_D})$ *and* $\widetilde{\mathbf{x}} := (y_{\widetilde{\mathbf{i}}_1}, \dots, y_{\widetilde{\mathbf{i}}_D})$ *is at most* $|\Omega| D / r$*.*

Finally, we rely on the following standard estimate of the expected distance of a one-dimensional random walk from the origin.

Proposition 6 (Expected distance traveled by one-dimensional random walk)**.** *Sample* $\mathbf{x} \in \{\pm 1\}^n$ *uniformly at random. Then* $\mathbb{E}\left[\,|\Sigma(\mathbf{x})|\,\right] = \Theta(\sqrt{n})$*.*

Now we are ready to prove the correlation bound.

Lemma 7 (Majority is moderately hard for shallow decision trees with respect to $\mu_{p,n}$)**.** *Let* $n, D \in \mathbb{N}$ *and* $p > 0$ *be such that* $(1-p)n$ *is an even positive integer. Let* $h \colon \{\pm 1\}^{pn} \to \{\pm 1\}$ *be a symmetric function. Then*

$$\mathsf{Corr}_{\mu_{p,n}}(h, \mathrm{DTDepth}[D]) \le \left| \mathop{\mathbb{E}}_{\mathbf{y} \sim \mu_{p,n}}[h(\mathbf{y})] \right| + O\left(\frac{1}{p} \cdot \sqrt{\frac{D}{n}} \right).$$

Proof. Let $T \colon \{\pm 1\}^{pn} \to \{\pm 1\}$ be a decision tree of depth at most D. Assume without loss of generality that $D \le pn$, T always makes precisely D queries, and T never queries the same variable twice.

Independently sample a string $\mathbf{y} \sim \mu_{p,n}$ and a uniform random permutation $\boldsymbol{\sigma} \colon [pn] \to [pn]$. For $y \in \{\pm 1\}^{pn}$, let $\boldsymbol{\sigma}(y)$ denote the string $(y_{\boldsymbol{\sigma}(1)}, \dots, y_{\boldsymbol{\sigma}(pn)})$. Observe that $\boldsymbol{\sigma}(\mathbf{y})$ is still distributed according to $\mu_{p,n}$, just like $\mathbf{y}$ itself. Therefore, instead of analyzing $T(\mathbf{y})$, it suffices to analyze $T(\boldsymbol{\sigma}(\mathbf{y}))$.

Let us first analyze $T(\boldsymbol{\sigma}(y))$ for a fixed $y \in \{\pm 1\}^{pn}$. We can imagine that the permutation $\boldsymbol{\sigma}$ is determined "on the fly," i.e., when T queries position i of its input, then the index $\boldsymbol{\sigma}(i)$ is chosen uniformly at random from among the unused indices. Thus, we have $T(\boldsymbol{\sigma}(y)) = f(y_{\mathbf{i}_1}, \dots, y_{\mathbf{i}_D})$, where $\mathbf{i}_1, \dots, \mathbf{i}_D \in [pn]$ are chosen uniformly at random without replacement and $f(z)$ is the label of the leaf reached by starting at the root of T and traversing the edges labeled $z_1, \dots, z_D$.

Sample $\widetilde{\mathbf{i}}_1, \dots, \widetilde{\mathbf{i}}_D \in [pn]$ uniformly at random *with* replacement. By Theorem 11, the total variation distance between the sequence $(y_{\mathbf{i}_1}, \dots, y_{\mathbf{i}_D})$ and the sequence $(y_{\widetilde{\mathbf{i}}_1}, \dots, y_{\widetilde{\mathbf{i}}_D})$ is at most $\frac{2D}{pn}$. Furthermore, sample $\mathbf{z} \in \{\pm 1\}^D$ uniformly at random; by Proposition 5, the total variation distance between $(y_{\widetilde{\mathbf{i}}_1}, \dots, y_{\widetilde{\mathbf{i}}_D})$ and $\mathbf{z}$ is at most $O(\frac{|\Sigma(y)|\sqrt{D}}{pn})$. Therefore,

$$\begin{aligned} \mathbb{E}[T(\boldsymbol{\sigma}(y)) \cdot h(\boldsymbol{\sigma}(y))] &= \mathbb{E}[f(y_{\mathbf{i}_1}, \dots, y_{\mathbf{i}_D}) \cdot h(y)] \\ &\le \mathbb{E}[f(\mathbf{z})] \cdot h(y) + O\left(\frac{D + |\Sigma(y)| \cdot \sqrt{D}}{pn} \right). \end{aligned}$$

Recalling now that $\mathbf{y} \sim \mu_{p,n}$ and $\mathbf{y}$ is independent of $\mathbf{z}$, we have

$$\begin{aligned}\mathsf{Corr}_{\mu_{p,n}}(h, T) &= \mathbb{E}[T(\boldsymbol{\sigma}(\mathbf{y})) \cdot h(\boldsymbol{\sigma}(\mathbf{y}))] \\ &\le \mathbb{E}[f(\mathbf{z})] \cdot \mathbb{E}[h(\mathbf{y})] + O\left(\frac{D + \mathbb{E}\left[|\Sigma(\mathbf{y})|\right] \cdot \sqrt{D}}{pn}\right) \\ &\le |\mathbb{E}[h(\mathbf{y})]| + O\left(\frac{D + \mathbb{E}\left[|\Sigma(\mathbf{y})|\right] \cdot \sqrt{D}}{pn}\right).\end{aligned}$$

Furthermore,

$$\mathbb{E}\left[|\Sigma(\mathbf{y})|\right] = \mathop{\mathbb{E}}_{\mathbf{x} \in \{\pm 1\}^n \setminus E_{p,n}}\left[|\Sigma(\mathbf{x})|\right] \le \mathop{\mathbb{E}}_{\mathbf{x} \in \{\pm 1\}^n}\left[|\Sigma(\mathbf{x})|\right] \le O(\sqrt{n})$$

by Proposition 6. Therefore,

$$\mathsf{Corr}_{\mu_{p,n}}(h, T) \le |\mathbb{E}[h(\mathbf{y})]| + O\left(\frac{D + \sqrt{Dn}}{pn}\right) = |\mathbb{E}[h(\mathbf{y})]| + O\left(\frac{1}{p} \cdot \sqrt{\frac{D}{n}}\right). \qquad \square$$

Corollary 1 (Majority retains structure under $\widetilde{\mathcal{R}}_{p,n}$)**.** *Let $n \in \mathbb{N}$ and $p > 0$ be such that $(1-p)n$ is an even positive integer. Let $h \colon \{\pm 1\}^n \to \{\pm 1\}$ be a symmetric function. If we sample $\boldsymbol{\pi} \sim \widetilde{\mathcal{R}}_{p,n}$, then except with probability $2\exp(-p^2 n/2)$, for every $D \in \mathbb{N}$, we have*

$$\mathsf{Corr}_{\mu_{p,n}}(h|_{\boldsymbol{\pi}}, \mathrm{DTDepth}[D]) \le \left|\mathop{\mathbb{E}}_{\mathbf{x} \in \{\pm 1\}^n}[h(\mathbf{x})]\right| + O\left(\frac{1}{p} \cdot \sqrt{\frac{D}{n}}\right).$$

Proof. Since h is symmetric, there is some function $f \colon \mathbb{Z} \to \{\pm 1\}$ such that $h(x) = f(\Sigma(x))$ for every x. By Hoeffding's inequality, $|E_{p,n}|/2^n \le 2\exp(-p^2 n/2)$. Whenever $\boldsymbol{\pi} \in R_{p,n}$, we have $h|_{\boldsymbol{\pi}}(y) = f(\Sigma(y))$, since $\boldsymbol{\pi}$ is balanced. In this case, Lemma 7 gives us

$$\mathsf{Corr}_{\mu_{p,n}}(h|_{\boldsymbol{\pi}}, \mathrm{DTDepth}[D]) \le \left|\mathop{\mathbb{E}}_{\mathbf{y} \sim \mu_{p,n}}[f(\Sigma(\mathbf{y}))]\right| + O\left(\frac{1}{p} \cdot \sqrt{\frac{D}{n}}\right). \tag{3}$$

If we sample $\mathbf{x} \in \{\pm 1\}^n$ uniformly at random, then the total variation distance between $\Sigma(\mathbf{y})$ and $\Sigma(\mathbf{x})$ is at most $|E_{p,n}|/2^n$. Therefore, (3) implies

$$\begin{aligned}\mathsf{Corr}_{\mu_{p,n}}(h|_{\boldsymbol{\pi}}, \mathrm{DTDepth}[D]) &\le |\mathbb{E}[f(\Sigma(\mathbf{x})]| + 4\exp(-p^2 n/2) + O\left(\frac{1}{p} \cdot \sqrt{\frac{D}{n}}\right) \\ &= |\mathbb{E}[h(\mathbf{x})]| + O\left(\frac{1}{p} \cdot \sqrt{\frac{D}{n}}\right). \qquad \square\end{aligned}$$

For clarity, let us see how our analysis so far implies a correlation bound for majority, i.e., the $t = 1$ case of Theorem 10.

Theorem 12 (Majority is moderately hard for AC^0_d circuits)**.** *Let $h \colon \{\pm 1\}^n \to \{\pm 1\}$ be a symmetric function, and let $g \colon \{\pm 1\}^n \to \{\pm 1\}$ be an AC^0_d circuit of size S.*

$$\mathsf{Corr}(g, h) \le \left|\mathop{\mathbb{E}}_{\mathbf{x} \in \{\pm 1\}^n}[h(\mathbf{x})]\right| + \frac{O(\log S)^{d-1} \cdot \sqrt{\log n}}{\sqrt{n}}.$$

Proof. Let $\zeta = |\mathbb{E}_{\mathbf{x}}[h(\mathbf{x})]|$. Let $p = 1/O(\log S)^{d-1}$ be the value from Lemma 5. Let D be a parameter that we will choose later. Let $\delta = O(n) \cdot 2^{-D}$ be the failure probability from Lemma 5 (with $t = 1$), let $\varepsilon = \zeta + O(\frac{1}{p} \cdot \sqrt{\frac{D}{n}})$ be the correlation bound from Corollary 1, and let $\gamma = 2\exp(-p^2 n/2) \ll \varepsilon$ be the failure probability from Corollary 1. Sample $\boldsymbol{\pi} \sim \widetilde{\mathcal{R}}_{p,n}$. Let $\mathbf{T} = g|_{\boldsymbol{\pi}}$ if $\mathrm{DTDepth}(g|_{\boldsymbol{\pi}}) \leq D$ and let $\mathbf{T}$ be a trivial decision tree otherwise. Then

$$\begin{aligned}
\mathsf{Corr}(h,g) &= \mathop{\mathbb{E}}_{\boldsymbol{\pi}}[\mathsf{Corr}_{\mu_{p,n}}(h|_{\boldsymbol{\pi}}, g|_{\boldsymbol{\pi}})] && \text{(Lemma 6)}\\
&\leq \delta + \mathop{\mathbb{E}}_{\boldsymbol{\pi}}[\mathsf{Corr}_{\mu_{p,n}}(h|_{\boldsymbol{\pi}}, \mathbf{T})] && \text{(Lemma 5)}\\
&\leq \delta + \mathop{\mathbb{E}}_{\boldsymbol{\pi}}\left[\mathsf{Corr}_{\mu_{p,n}}(h|_{\boldsymbol{\pi}}, \mathrm{DTDepth}[D])\right]\\
&\leq \delta + \gamma + \zeta + \varepsilon && \text{(Corollary 1)}\\
&= \zeta + O(n) \cdot 2^{-D} + \frac{O(\log S)^{d-1} \cdot \sqrt{D}}{\sqrt{n}}.
\end{aligned}$$

Pick $D = O(\log n)$ to complete the proof. □

5.2 Improved XOR Lemma for the Random Simplification Method

To amplify the hardness of majority, we will present a more sophisticated version of our XOR lemma for the random simplification method, based on our XOR lemma for decision trees (Lemma 3). We use the following standard fact.

Proposition 7 (Concave $\implies$ subadditive)**.** *Let $f \colon [0,\infty) \to (0,\infty)$ be a log-concave function, and assume that $f(0) \geq 1$. Then $f(x+y) \leq f(x) \cdot f(y)$ for every $x, y \in [0,\infty)$.*

Proof. First, suppose $x = y = 0$. Then $f(x) \cdot f(y) = f(0)^2 \geq f(0) = f(x+y)$ because $f(0) \geq 1$. Now, suppose $x + y > 0$. Since f is log-concave, for any $z, \lambda \in [0,\infty)$, we have

$$f(\lambda z) = f(\lambda z + (1-\lambda)0) \geq f(x)^{\lambda} f(0)^{1-\lambda} \geq f(z)^{\lambda}.$$

Therefore, letting $z = x + y > 0$ and $\lambda = x/z$, we have

$$f(x) \cdot f(y) = f(\lambda z) \cdot f((1-\lambda)z) \geq f(z)^{\lambda} \cdot f(z)^{1-\lambda} = f(z).$$ □

We also rely on the following trivial fact, which says that computing an XOR of functions can only get more difficult if we introduce more functions into the XOR.

Proposition 8 (Computing the XOR of more functions is harder)**.** *Let $h_1, \dots, h_t \colon \{\pm 1\}^r \to \{\pm 1\}$ be functions, let $\mu_1, \dots, \mu_t$ be distributions over $\{\pm 1\}^r$, and let $I \subseteq [t]$. Define $h \colon \{\pm 1\}^{rt} \to \{\pm 1\}$ by $h(y^{(1)}, \dots, y^{(t)}) = \prod_{i=1}^{t} h_i(y^{(i)})$, and define $h_I \colon \{\pm 1\}^{r|I|} \to \{\pm 1\}$ by $h((y^{(i)})_{i \in I}) = \prod_{i \in I} h_i(y^{(i)})$. Let $\mu = \mu_1 \otimes \cdots \otimes \mu_t$, and let $\mu_I = \bigotimes_{i \in I} \mu_i$. Then for any $D \in \mathbb{N}$, we have*

$$\mathsf{Corr}_{\mu}(h, \mathrm{DTDepth}[D]) \leq \mathsf{Corr}_{\mu_I}(h_I, \mathrm{DTDepth}[D]).$$

Proof. Let $T \colon \{\pm 1\}^{rt} \to \{\pm 1\}$ be any decision tree of depth D. Let $J = [t] \setminus I$. For simplicity, if $\vec{y}_I \in \{\pm 1\}^{r|I|}$ and $\vec{y}_J \in \{\pm 1\}^{r|J|}$, we will write $(\vec{y}_I, \vec{y}_J)$ to denote the string in $\{\pm 1\}^{rt}$ obtained by using $\vec{y}_I$ to fill in the blocks in I and using $\vec{y}_J$ to fill in the blocks in J. Then

$$\begin{aligned}
\mathsf{Corr}_\mu(h, T) &= \mathop{\mathbb{E}}_{\vec{\mathbf{y}} \sim \mu}[h(\vec{\mathbf{y}}) \cdot T(\vec{\mathbf{y}})] \\
&= \mathop{\mathbb{E}}_{\vec{\mathbf{y}}_J \sim \mu_J}\left[h_J(\vec{\mathbf{y}}_J) \cdot \mathop{\mathbb{E}}_{\vec{\mathbf{y}}_I \sim \mu_I}[h_I(\vec{\mathbf{y}}_I) \cdot T(\vec{\mathbf{y}}_I, \vec{\mathbf{y}}_J)]\right] \\
&\le \mathop{\mathbb{E}}_{\vec{\mathbf{y}}_J \sim \mu_J}\left[\left|\mathop{\mathbb{E}}_{\vec{\mathbf{y}}_I \sim \mu_I}[h_I(\vec{\mathbf{y}}_I) \cdot T(\vec{\mathbf{y}}_I, \vec{\mathbf{y}}_J)]\right|\right] \\
&\le \mathop{\mathbb{E}}_{\vec{\mathbf{y}}_J \sim \mu_J}\left[\mathsf{Corr}_{\mu_I}(h_I, \mathrm{DTDepth}[D])\right],
\end{aligned}$$

because depth-D decision trees are closed under restricting variables and complementing the output value. □

Now we can state and prove our improved XOR lemma for the random simplification method.

Lemma 8 (Tighter XOR lemma for the random simplification method)**.** *Let $n, t \in \mathbb{N}$ and let $h\colon \{\pm 1\}^n \to \{\pm 1\}$ and $g\colon \{\pm 1\}^{nt} \to \{\pm 1\}$ be Boolean functions. Let $r \in \mathbb{N}$, let $\mathcal{R}$ be a distribution over generalized restrictions $\boldsymbol{\pi}\colon \{\pm 1\}^r \to \{\pm 1\}^n$, and let μ be a distribution over $\{\pm 1\}^r$. Let $\varepsilon, \delta\colon [0, \infty) \to (0, \infty)$ and $\gamma > 0$. Assume the following.*

1. *If we sample $\boldsymbol{\pi} \sim \mathcal{R}$ and $\mathbf{y} \sim \mu$ independently, then $\boldsymbol{\pi}(\mathbf{y})$ is distributed uniformly over $\{\pm 1\}^n$.*
2. *For every $D \in \mathbb{N}$, we have*
$$\Pr_{\vec{\boldsymbol{\pi}} \sim \mathcal{R}^{\otimes t}}[\mathrm{DTDepth}(g|_{\vec{\boldsymbol{\pi}}}) > D] \le \delta(D).$$
3. *With probability $1-\gamma$ over $\boldsymbol{\pi} \sim \mathcal{R}$, for every $D \in \mathbb{N}$, we have $\mathsf{Corr}_\mu(h|_{\boldsymbol{\pi}}, \mathrm{DTDepth}_\mu[D]) \le \varepsilon(D)$.*
4. *The functions $\varepsilon(D)$ and $\delta(D)$ are log-concave, and $\delta(0) \ge 1$.*

Then for every integer $D \ge 2$, we have

$$\mathsf{Corr}(g, h^{\oplus t}) \le O\Big(\sqrt{\gamma} + \varepsilon(4D) + \delta(D)\Big)^t.$$

Proof. Sample $\vec{\boldsymbol{\pi}} = (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_t) \sim \mathcal{R}^{\otimes t}$. Let $\mathbf{I}$ be the set of $i \in [t]$ such that $h|_{\boldsymbol{\pi}_i}$ is hard for decision trees as described in Assumption 3, i.e., $i \in \mathbf{I}$ if and only if for every $D \in \mathbb{N}$, we have $\mathsf{Corr}(h|_{\boldsymbol{\pi}_i}, \mathrm{DTDepth}[D]) \le \varepsilon(D)$. Identify $\mathbf{I}$ with its indicator function, i.e., $\mathbf{I}(i) = 1 \iff i \in \mathbf{I}$. We define a decision tree $\mathbf{T}$ as follows.

- If $\mathrm{DTDepth}(g|_{\vec{\boldsymbol{\pi}}}) \le Dt$ and $|\mathbf{I}| \ge t/2$, then we let $\mathbf{T} = g|_{\vec{\boldsymbol{\pi}}}$.
- Otherwise, we let $\mathbf{T}$ be the constant one function (a depth-zero tree).

Observe that the second case occurs with probability at most $\delta(Dt) + 2^t \gamma^{t/2}$. Furthermore, in either case, the depth of $\mathbf{T}$ is at most $2D \cdot |\mathbf{I}|$. Define $h_{\mathbf{I}}\colon \{\pm 1\}^{r|\mathbf{I}|} \to \{\pm 1\}$ by

$$h_{\mathbf{I}}((y^{(i)})_{i \in \mathbf{I}}) = \prod_{i \in \mathbf{I}} h|_{\boldsymbol{\pi}_i}(y^{(i)}).$$

Then

$$\begin{aligned}
\mathsf{Corr}(h^{\oplus t}, g) &= \mathop{\mathbb{E}}_{\vec{\pi}}[\mathsf{Corr}_{\mu^{\otimes t}}(h^{\oplus t}|_{\vec{\pi}}, g|_{\vec{\pi}})] \\
&\leq \delta(Dt) + 2^t\gamma^{t/2} + \mathop{\mathbb{E}}_{\vec{\pi}}\left[\mathsf{Corr}_{\mu^{\otimes t}}(h^{\oplus t}|_{\vec{\pi}}, \mathbf{T})\right] \\
&\leq \delta(Dt) + 2^t\gamma^{t/2} + \mathop{\mathbb{E}}_{\vec{\pi}}\left[\mathsf{Corr}_{\mu^{\otimes t}}\left(h^{\oplus t}|_{\vec{\pi}}, \mathrm{DTDepth}\left[2D \cdot |\mathbf{I}|\right]\right)\right] \\
&\leq \delta(Dt) + 2^t\gamma^{t/2} + \mathop{\mathbb{E}}_{\vec{\pi}}\left[\mathsf{Corr}_{\mu^{\otimes |\mathbf{I}|}}\left(h_{\mathbf{I}}, \mathrm{DTDepth}\left[2D \cdot |\mathbf{I}|\right]\right)\right] && \text{(Proposition 8)} \\
&\leq \delta(Dt) + 2^t\gamma^{t/2} + \mathop{\mathbb{E}}_{\vec{\pi}}\left[O(\varepsilon(4D))^{|\mathbf{I}|}\right] && \text{(Lemma 3)} \\
&= \delta(Dt) + 2^t\gamma^{t/2} + \mathop{\mathbb{E}}_{\vec{\pi}}\left[\prod_{i=1}^{t} O(\varepsilon(4D))^{\mathbf{I}(i)}\right] \\
&= \delta(Dt) + 2^t\gamma^{t/2} + \prod_{i=1}^{t} \mathop{\mathbb{E}}_{\vec{\pi}}[O(\varepsilon(4D))^{\mathbf{I}(i)}] && \text{(Independence)} \\
&\leq \delta(Dt) + 2^t\gamma^{t/2} + (\gamma + O(\varepsilon(4D)))^t.
\end{aligned}$$

By Proposition 7, we have $\delta(Dt) \leq \delta(D)^t$. Simplifying, we get a bound of

$$\delta(D)^t + (2\sqrt{\gamma})^t + (\gamma + O(\varepsilon(4D))^t \leq O\Big(\sqrt{\gamma} + \varepsilon(4D) + \delta(D)\Big)^t. \qquad \square$$

Combining this XOR lemma with our random-restrictions-based proof that majority is average-case-hard for AC^0_d circuits will complete the proof of Theorem 10:

Proof of Theorem 10. Let $\zeta = |\mathbb{E}_{\mathbf{x} \in \{\pm 1\}^n}[h(\mathbf{x})]|$. By Lemma 5 and Corollary 1, the assumptions of Lemma 8 are satisfied with $\mathcal{R} = \widetilde{\mathcal{R}}_{p,n}$, $p = 1/O(\log S)^{d-1}$, $\mu = \mu_{p,n}$, $\delta(D) = O(n) \cdot 2^{-D}$, $\varepsilon(D) = \zeta + O(\frac{1}{p\sqrt{n}}) \cdot \sqrt{D}$, and $\gamma = 2\exp(-2p^2 n) \ll \frac{1}{p^2 n}$. Note that $\delta(D)$ and $\varepsilon(D)$ are both log-concave and $\delta(0) \geq 1$. Therefore, Lemma 8 gives us a correlation bound of

$$O\left(\zeta + \frac{\sqrt{D}}{p \cdot \sqrt{n}} + n \cdot 2^{-D}\right)^t.$$

Choosing $D = O(\log n)$ completes the proof. $\square$

6 Directions for Further Research

The main open question related to our work is whether XORing always amplifies hardness for AC^0 circuits (cf. Question 1). We wish to also highlight the problem of proving *tight* correlation bounds for depth reduction within AC^0 (cf. Theorem 2). That is, what is the correlation between linear-size AC^0_{d+k} circuits and near-exponential-size AC^0_d circuits?

For simplicity, let us consider the case that d and k are both constants. As discussed previously, the extreme case $k = 1$ (i.e., using AC^0_d circuits to approximate AC^0_{d+1} circuits) is resolved by Håstad, Rossman, Servedio, and Tan's work [HRST17] to within polynomial factors; the optimal correlation bound is $n^{\Theta(1)}$. Prior work also implies near-matching upper and lower bounds in the opposite extreme case $d = 1$ (i.e., using AC^0_1 circuits to approximate AC^0_{1+k} circuits). In this case, it turns out that the optimal correlation bound is $\exp\left(-\widetilde{\Theta}(\log^k n)\right)$. (The approximators are based on the Linial-Nisan-Mansour theorem [LMN93]; see Appendix B for details.)

Based on those two extreme cases, it is tempting to conjecture that for all d and k, the optimal correlation bound should be $\exp\left(-\widetilde{\Theta}(\log^k n)\right)$, but in truth it is not at all clear that this is the best guess. Arguably the most interesting case is $k = 2$, i.e., the problem of using AC^0_d circuits to approximate AC^0_{d+2} circuits. On the one hand, the best method we know for constructing such an approximator is simply to use an optimal AC^0_1 approximator. On the other hand, the best correlation bound we know for this case is Håstad, Rossman, Servedio, and Tan's bound [HRST17]. We therefore have a considerable gap between the upper and lower correlation bounds for this case, namely $n^{-\Omega(1)}$ vs. $n^{-\widetilde{O}(\log^d n)}$.

Acknowledgments

I thank Avishay Tal for collaboration at an early stage of this project. I thank Li-Yang Tan for encouragement. I thank Pooya Hatami for a helpful conversation.

References

[AASY16] Benny Applebaum, Sergei Artemenko, Ronen Shaltiel, and Guang Yang. "Incompressible functions, relative-error extractors, and the power of nondeterministic reductions". In: *Comput. Complexity* 25.2 (2016), pp. 349–418. ISSN: 1016-3328. DOI: 10.1007/s00037-016-0128-9.

[ABFR94] J. Aspnes, R. Beigel, M. Furst, and S. Rudich. "The expressive power of voting polynomials". In: *Combinatorica* 14.2 (1994), pp. 135–148. ISSN: 0209-9683. DOI: 10.1007/BF01215346.

[AG94] Eric Allender and Vivek Gore. "A Uniform Circuit Lower Bound for the Permanent". In: *SIAM Journal on Computing* 23.5 (1994), pp. 1026–1049. DOI: 10.1137/S0097539792233907.

[AGM03] Noga Alon, Oded Goldreich, and Yishay Mansour. "Almost k-wise independence versus k-wise independence". In: *Inform. Process. Lett.* 88.3 (2003), pp. 107–110. ISSN: 0020-0190. DOI: 10.1016/S0020-0190(03)00359-4.

[AH94] Eric Allender and Ulrich Hertrampf. "Depth reduction for circuits of unbounded fan-in". In: *Inform. and Comput.* 112.2 (1994), pp. 217–238. ISSN: 0890-5401. DOI: 10.1006/inco.1994.1057.

[Ajt83] M. Ajtai. "Σ^1_1-formulae on finite structures". In: *Ann. Pure Appl. Logic* 24.1 (1983), pp. 1–48. DOI: 10.1016/0168-0072(83)90038-6.

[All89] Eric Allender. "A note on the power of threshold circuits". In: *Proceedings of the 30th Symposium on Foundations of Computer Science (FOCS)*. 1989, pp. 580–584. DOI: 10.1109/SFCS.1989.63538.

[Ama09] Kazuyuki Amano. "Bounds on the size of small depth circuits for approximating majority". In: *Proceedings of the 36th International Colloquium on Automata, Languages and Programming (ICALP)*. 2009, pp. 59–70. DOI: 10.1007/978-3-642-02927-1_7.

[AŠW09] Andris Ambainis, Robert Špalek, and Ronald de Wolf. "A new quantum lower bound method, with applications to direct product theorems and time-space tradeoffs". In: *Algorithmica* 55.3 (2009), pp. 422–461. ISSN: 0178-4617. DOI: 10.1007/s00453-007-9022-9.

[Bab87] László Babai. "Random oracles separate PSPACE from the polynomial-time hierarchy". In: *Inform. Process. Lett.* 26.1 (1987), pp. 51–53. DOI: 10.1016/0020-0190(87)90036-6.

[BAN95] Y. Ben-Asher and I. Newman. "Decision trees with AND, OR queries". In: *Proceedings of the 10th Conference on Structure in Complexity Theory (SCT)*. 1995, pp. 74–81. DOI: 10.1109/SCT.1995.514729.

[Baz09] Louay M. J. Bazzi. "Polylogarithmic Independence Can Fool DNF Formulas". In: *SIAM J. Comput.* 38.6 (2009), pp. 2220–2272. DOI: 10.1137/070691954.

[BB19] Eric Blais and Joshua Brody. "Optimal Separation and Strong Direct Sum for Randomized Query Complexity". In: *Proceedings of the 34th Computational Complexity Conference (CCC)*. 2019, 29:1–29:17. DOI: 10.4230/LIPIcs.CCC.2019.29.

[BIS12] Paul Beame, Russell Impagliazzo, and Srikanth Srinivasan. "Approximating AC^0 by small height decision trees and a deterministic algorithm for $\#\mathrm{AC}^0 SAT$". In: *27th Conference on Computational Complexity (CCC)*. 2012, pp. 117–125. DOI: 10.1109/CCC.2012.40.

[BK18] Shalev Ben-David and Robin Kothari. "Randomized query complexity of sabotaged and composed functions". In: *Theory Comput.* 14 (2018), Paper No. 5, 27. DOI: 10.4086/toc.2018.v014a005.

[BKLS20] Joshua Brody, Jae Tak Kim, Peem Lerdputtipongporn, and Hariharan Srinivasulu. *A Strong XOR Lemma for Randomized Query Complexity*. 2020. arXiv: 2007.05580.

[Bop97] Ravi B. Boppana. "The average sensitivity of bounded-depth circuits". In: *Information Processing Letters* 63.5 (1997), pp. 257–261. ISSN: 0020-0190. DOI: 10.1016/S0020-0190(97)00131-2.

[Bra10] Mark Braverman. "Polylogarithmic independence fools AC^0 circuits". In: *Journal of the ACM* 57.5 (2010).

[BRS91] Richard Beigel, Nick Reingold, and Daniel A. Spielman. "The perceptron strikes back". In: *Proceedings of the 6th Annual Structure in Complexity Theory Conference (SCT)*. 1991, pp. 286–291. DOI: 10.1109/SCT.1991.160270.

[BT94] Richard Beigel and Jun Tarui. "On ACC". In: *Comput. Complexity* 4.4 (1994), pp. 350–366. DOI: 10.1007/BF01263423.

[Che23] Lijie Chen. "New Lower Bounds and Derandomization for ACC, and a Derandomization-Centric View on the Algorithmic Method". In: *14th Innovations in Theoretical Computer Science Conference (ITCS)*. 2023, 34:1–34:15. DOI: 10.4230/LIPIcs.ITCS.2023.34.

[CHHLZ20] Eshan Chattopadhyay, Pooya Hatami, Kaave Hosseini, Shachar Lovett, and David Zuckerman. "XOR lemmas for resilient functions against polynomials". In: *Proceedings of the 52nd Symposium on Theory of Computing (STOC)*. 2020, pp. 234–246. DOI: 10.1145/3357713.3384242.

[CHLR23] Yeyuan Chen, Yizhi Huang, Jiatu Li, and Hanlin Ren. "Range avoidance, remote point, and hard partial truth table via satisfying-pairs algorithms". In: *Proceedings of the 55th Symposium on Theory of Computing (STOC)*. 2023, pp. 1058–1066. DOI: 10.1145/3564246.3585147.

[CL21] Lijie Chen and Xin Lyu. "Inverse-exponential correlation bounds and extremely rigid matrices from a new derandomized XOR lemma". In: *Proceedings of the 53rd Symposium on Theory of Computing (STOC)*. 2021, pp. 761–771. DOI: 10.1145/3406325.3451132.

[CLLO21] Lijie Chen, Zhenjian Lu, Xin Lyu, and Igor C. Oliveira. "Majority vs. approximate linear sum and average-case complexity below NC^1". In: *48th International Colloquium on Automata, Languages, and Programming (ICALP)*. 2021, 51:1–51:20. DOI: 10.4230/LIPIcs.ICALP.2021.51.

[CLW20] Lijie Chen, Xin Lyu, and R. Ryan Williams. "Almost-everywhere circuit lower bounds from non-trivial derandomization". In: *61st Symposium on Foundations of Computer Science (FOCS)*. 2020, pp. 1–12. DOI: 10.1109/FOCS46700.2020.00009.

[CP19] Shiteng Chen and Periklis A. Papakonstantinou. "Depth reduction for composites". In: *SIAM J. Comput.* 48.2 (2019), pp. 668–686. ISSN: 0097-5397. DOI: 10.1137/17M1129672.

[CR22] Lijie Chen and Hanlin Ren. "Strong Average-Case Circuit Lower Bounds from Nontrivial Derandomization". In: *SIAM Journal on Computing* 51.3 (2022), STOC20–115–STOC20–173. DOI: 10.1137/20M1364886.

[DF80] P. Diaconis and D. Freedman. "Finite exchangeable sequences". In: *Ann. Probab.* 8.4 (1980), pp. 745–764. ISSN: 0091-1798. DOI: 10.1214/aop/1176994663.

[Dru12] Andrew Drucker. "Improved direct product theorems for randomized query complexity". In: *Comput. Complexity* 21.2 (2012), pp. 197–244. ISSN: 1016-3328. DOI: 10.1007/s00037-012-0043-7.

[Fil10] Yuval Filmus. "Smolensky's Lower Bound". 2010. URL: https://yuvalfilmus.cs.technion.ac.il/Manuscripts/Smolensky.pdf.

[FSS84] Merrick Furst, James B. Saxe, and Michael Sipser. "Parity, circuits, and the polynomial-time hierarchy". In: *Math. Systems Theory* 17.1 (1984), pp. 13–27. DOI: 10.1007/BF01744431.

[GGHKR07] Shafi Goldwasser, Dan Gutfreund, Alexander Healy, Tali Kaufman, and Guy N. Rothblum. "Verifying and decoding in constant depth". In: *Proceedings of the 39th Symposium on Theory of Computing (STOC)*. 2007, pp. 440–449. DOI: 10.1145/1250790.1250855.

[GNW11] Oded Goldreich, Noam Nisan, and Avi Wigderson. "On Yao's XOR-lemma". In: *Studies in complexity and cryptography*. Vol. 6650. Lecture Notes in Comput. Sci. Springer, Heidelberg, 2011, pp. 273–301. DOI: 10.1007/978-3-642-22670-0_23.

[GR08] Dan Gutfreund and Guy N. Rothblum. "The complexity of local list decoding". In: *Proceedings of the 12th International Conference on Randomization and Computation (RANDOM)*. 2008, pp. 455–468. DOI: 10.1007/978-3-540-85363-3_36.

[GSV18] Aryeh Grinberg, Ronen Shaltiel, and Emanuele Viola. "Indistinguishability by adaptive procedures with advice, and lower bounds on hardness amplification proofs". In: *Proceedings of the 59th Symposium on Foundations of Computer Science (FOCS)*. 2018, pp. 956–966. DOI: 10.1109/FOCS.2018.00094.

[Hås01] Johan Håstad. "A slight sharpening of LMN". In: *Journal of Computer and System Sciences* 63.3 (2001), pp. 498–508. ISSN: 0022-0000. DOI: 10.1006/jcss.2001.1803.

[Hås14] Johan Håstad. "On the correlation of parity and small-depth circuits". In: *SIAM J. Comput.* 43.5 (2014), pp. 1699–1708. DOI: 10.1137/120897432.

[Hås86a] Johan Håstad. "Almost Optimal Lower Bounds for Small Depth Circuits". In: *Proceedings of the 18th Symposium on Theory of Computing (STOC)*. 1986, 6–20. DOI: 10.1145/12130.12132.

[Hås86b] Johan Håstad. "Computational limitations for small depth circuits". PhD thesis. Massachusetts Institute of Technology, 1986.

[HHTT23] Pooya Hatami, William M. Hoza, Avishay Tal, and Roei Tell. "Depth-d Threshold Circuits vs. Depth-$(d+1)$ AND-OR Trees". In: *Proceedings of the 55th Symposium on Theory of Computing (STOC)*. Full version: https://eccc.weizmann.ac.il/report/2022/087/. 2023, 895–904. DOI: 10.1145/3564246.3585216.

[HMPST93] András Hajnal, Wolfgang Maass, Pavel Pudlák, Mario Szegedy, and György Turán. "Threshold Circuits of Bounded Depth". In: *Journal of Computer and System Sciences* 46.2 (1993), pp. 129–154. DOI: 10.1016/0022-0000(93)90001-D.

[HRST17] Johan Håstad, Benjamin Rossman, Rocco A. Servedio, and Li-Yang Tan. "An average-case depth hierarchy theorem for Boolean circuits". In: *J. ACM* 64.5 (2017), Art. 35, 27. ISSN: 0004-5411. DOI: 10.1145/3095799.

[HS19] Prahladh Harsha and Srikanth Srinivasan. "On polynomial approximations to AC^0". In: *Random Structures Algorithms* 54.2 (2019), pp. 289–303. DOI: 10.1002/rsa.20786.

[HV21] Xuangui Huang and Emanuele Viola. "Average-case rigidity lower bounds". In: *Proceedings of the 16th International Computer Science Symposium in Russia (CSR)*. 2021, pp. 186–205. DOI: 10.1007/978-3-030-79416-3_11.

[IMP12] Russell Impagliazzo, William Matthews, and Ramamohan Paturi. "A satisfiability algorithm for AC^0". In: *Proceedings of the 23rd Symposium on Discrete Algorithms (SODA)*. 2012, pp. 961–972. DOI: 10.1137/1.9781611973099.77.

[Imp95] Russell Impagliazzo. "Hard-core distributions for somewhat hard problems". In: *36th Symposium on Foundations of Computer Science (FOCS)*. 1995, pp. 538–545. DOI: 10.1109/SFCS.1995.492584.

[IRW94] Russell Impagliazzo, Ran Raz, and Avi Wigderson. "A direct product theorem". In: *Proceedings of 9th Annual Conference on Structure in Complexity Theory (SCT)*. 1994, pp. 88–96. DOI: 10.1109/SCT.1994.315814.

[JKS10] Rahul Jain, Hartmut Klauck, and Miklos Santha. "Optimal direct sum results for deterministic and randomized decision tree complexity". In: *Inform. Process. Lett.* 110.20 (2010), pp. 893–897. ISSN: 0020-0190. DOI: 10.1016/j.ipl.2010.07.020.

[KKL88] Jeff Kahn, Gil Kalai, and Nathan Linial. "The influence of variables on Boolean functions". In: *Proceedings of the 29th Symposium on Foundations of Computer Science (FOCS)*. 1988, pp. 68–80. DOI: 10.1109/SFCS.1988.21923.

[Kli01] Adam R. Klivans. “On the derandomization of constant depth circuits”. In: *Proceedings of the 5th International Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM)*. 2001, pp. 249–260. DOI: 10.1007/3-540-44666-4_28.

[KMS12] Jeff Kinne, Dieter van Melkebeek, and Ronen Shaltiel. “Pseudorandom generators, typically-correct derandomization, and circuit lower bounds”. In: *Comput. Complexity* 21.1 (2012), pp. 3–61. ISSN: 1016-3328. DOI: 10.1007/s00037-011-0019-z.

[KS18] Swastik Kopparty and Srikanth Srinivasan. “Certifying Polynomials for $\mathrm{AC}^0[\oplus]$ Circuits, with Applications to Lower Bounds and Circuit Compression”. In: *Theory of Computing* 14.12 (2018), pp. 1–24. DOI: 10.4086/toc.2018.v014a012.

[KŠW07] Hartmut Klauck, Robert Špalek, and Ronald de Wolf. “Quantum and classical strong direct product theorems and optimal time-space tradeoffs”. In: *SIAM J. Comput.* 36.5 (2007), pp. 1472–1493. ISSN: 0097-5397. DOI: 10.1137/05063235X.

[Lev87] L. A. Levin. “One way functions and pseudorandom generators”. In: *Combinatorica* 7.4 (1987), pp. 357–363. ISSN: 0209-9683. DOI: 10.1007/BF02579323.

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. “Constant depth circuits, Fourier transform, and learnability”. In: *Journal of the ACM* 40.3 (1993), pp. 607–620. DOI: 10.1145/174130.174138.

[LR13] Troy Lee and Jérémie Roland. “A strong direct product theorem for quantum query complexity”. In: *Comput. Complexity* 22.2 (2013), pp. 429–462. ISSN: 1016-3328. DOI: 10.1007/s00037-013-0066-8.

[LZ19] Fu Li and David Zuckerman. “Improved extractors for recognizable and algebraic sources”. In: *Proceedings of the 23rd International Conference on Randomization and Computation (RANDOM)*. 2019, 72:1–72:22. DOI: 10.4230/LIPIcs.APPROX-RANDOM.2019.72.

[MST06] Elchanan Mossel, Amir Shpilka, and Luca Trevisan. “On ε-biased generators in NC^0”. In: *Random Structures Algorithms* 29.1 (2006), pp. 56–81. ISSN: 1042-9832. DOI: 10.1002/rsa.20112.

[Nis91] Noam Nisan. “Pseudorandom bits for constant depth circuits”. In: *Combinatorica* 11.1 (1991), pp. 63–70. ISSN: 0209-9683. DOI: 10.1007/BF01375474.

[NRS99] Noam Nisan, Steven Rudich, and Michael Saks. “Products and help bits in decision trees”. In: *SIAM J. Comput.* 28.3 (1999), pp. 1035–1050. ISSN: 0097-5397. DOI: 10.1137/S0097539795282444.

[NW94] Noam Nisan and Avi Wigderson. “Hardness vs. randomness”. In: *J. Comput. System Sci.* 49.2 (1994), pp. 149–167. ISSN: 0022-0000. DOI: 10.1016/S0022-0000(05)80043-1.

[OSS19] Igor Carboni Oliveira, Rahul Santhanam, and Srikanth Srinivasan. “Parity helps to compute majority”. In: *34th Computational Complexity Conference (CCC)*. 2019, 23:1–23:17. DOI: 10.4230/LIPIcs.CCC.2019.23.

[OW07] Ryan O’Donnell and Karl Wimmer. “Approximation by DNF: examples and counterexamples”. In: *Proceedings of the 34th International Colloquium on Automata, Languages and Programming (ICALP)*. 2007, pp. 195–206. DOI: 10.1007/978-3-540-73420-8_19.

[Raz09] Alexander Razborov. "A Simple Proof of Bazzi's Theorem". In: *ACM Transactions on Computation Theory* 1.1 (2009). DOI: 10.1145/1490270.1490273.

[Raz87] Alexander A. Razborov. "Lower bounds on the size of constant-depth networks over a complete basis with logical addition". In: *Mathematical Notes of the Academy of Science of the USSR* 41.4 (1987), pp. 333–338. DOI: 10.1007/BF01137685.

[Ros17] Benjamin Rossman. "An entropy proof of the switching lemma and tight bounds on the decision-tree size of AC^0". 2017. URL: https://users.cs.duke.edu/~br148/logsize.pdf.

[RS19] Benjamin Rossman and Srikanth Srinivasan. "Separation of $\mathsf{AC}^0[\oplus]$ formulas and circuits". In: *Theory Comput.* 15 (2019), Paper No. 17, 20. DOI: 10.4086/toc.2019.v015a017.

[Sha03] Ronen Shaltiel. "Towards proving strong direct product theorems". In: *Comput. Complexity* 12.1-2 (2003), pp. 1–22. ISSN: 1016-3328. DOI: 10.1007/s00037-003-0175-x.

[Sha11] Ronen Shaltiel. "Weak derandomization of weak algorithms: explicit versions of Yao's lemma". In: *Comput. Complexity* 20.1 (2011), pp. 87–143. DOI: 10.1007/s00037-011-0006-4.

[Sha23] Ronen Shaltiel. "Is it possible to improve Yao's XOR lemma using reductions that exploit the efficiency of their oracle?" In: *Comput. Complexity* 32.1 (2023), Paper No. 5, 47. ISSN: 1016-3328. DOI: 10.1007/s00037-023-00238-9.

[She12] Alexander A. Sherstov. "Strong direct product theorems for quantum communication and query complexity". In: *SIAM J. Comput.* 41.5 (2012), pp. 1122–1165. ISSN: 0097-5397. DOI: 10.1137/110842661.

[Sip83] Michael Sipser. "Borel Sets and Circuit Complexity". In: *Proceedings of the 15th Symposium on Theory of Computing*. 1983, pp. 61–69. DOI: 10.1145/800061.808733.

[Smo87] Roman Smolensky. "Algebraic methods in the theory of lower bounds for Boolean circuit complexity". In: *Proceedings of the 19th Symposium on Theory of Computing (STOC)*. 1987, pp. 77–82. DOI: 10.1145/28395.28404.

[Smo93] Roman Smolensky. "On representations by low-degree polynomials". In: *Proceedings of 34th Annual Symposium on Foundations of Computer Science (FOCS)*. 1993, pp. 130–138. DOI: 10.1109/SFCS.1993.366874.

[Špa08] Robert Špalek. "The multiplicative quantum adversary". In: *Proceedings of the 23rd Conference on Computational Complexity (CCC)*. 2008, pp. 237–248. DOI: 10.1109/CCC.2008.9.

[SV10] Ronen Shaltiel and Emanuele Viola. "Hardness amplification proofs require majority". In: *SIAM J. Comput.* 39.7 (2010), pp. 3122–3154. ISSN: 0097-5397. DOI: 10.1137/080735096.

[Sze89] Mario Szegedy. "Algebraic methods in lower bounds for computational models with limited communication". PhD thesis. University of Chicago, 1989.

[Tal17] Avishay Tal. "Tight Bounds on the Fourier Spectrum of AC0". In: *Proceedings of the 32nd Computational Complexity Conference (CCC)*. 2017, 15:1–15:31. DOI: 10.4230/LIPIcs.CCC.2017.15.

[Tar93] Jun Tarui. “Probabilistic Polynomials, AC^0 Functions and the Polynomial-time Hierarchy”. In: *Theoretical Computer Science* 113.1 (1993), pp. 167–183. DOI: 10.1016/0304-3975(93)90214-E.

[Tel20] Roei Tell. *On implications of better sub-exponential lower bounds for $\mathcal{AC}^0$*. https://sites.google.com/site/roeitell/Expositions. 2020.

[Tod91] Seinosuke Toda. “PP is as hard as the polynomial-time hierarchy”. In: *SIAM J. Comput.* 20.5 (1991), pp. 865–877. ISSN: 0097-5397. DOI: 10.1137/0220053.

[Val77] Leslie G. Valiant. “Graph-theoretic arguments in low-level complexity”. In: *Proceedings of the 6th Symposium on Mathematical Foundations of Computer Science (MFCS)*. 1977, pp. 162–176. DOI: 10.1007/3-540-08353-7_135.

[Vio06] Emanuele Viola. “The complexity of hardness amplification and derandomization”. PhD thesis. Harvard University, 2006.

[Vio09] Emanuele Viola. “On the power of small-depth computation”. In: *Found. Trends Theor. Comput. Sci.* 5.1 (2009), pp. 1–72. ISSN: 1551-305X. DOI: 10.1561/0400000033.

[Vio12] Emanuele Viola. “The complexity of distributions”. In: *SIAM J. Comput.* 41.1 (2012), pp. 191–218. ISSN: 0097-5397. DOI: 10.1137/100814998.

[Vio17] Emanuele Viola. “Selected Challenges in Computational Lower Bounds”. In: *SIGACT News* 48.1 (2017), 39–45. ISSN: 0163-5700. DOI: 10.1145/3061640.3061648.

[Vio19] Emanuele Viola. *Matching Smolensky’s correlation bound with majority*. https://eccc.weizmann.ac.il/report/2019/175/. 2019.

[Vio20] Emanuele Viola. *New lower bounds for probabilistic degree and AC0 with parity gates*. https://eccc.weizmann.ac.il/report/2020/015/. 2020.

[Wil14] Ryan Williams. “Nonuniform ACC Circuit Lower Bounds”. In: *J. ACM* 61.1 (2014). ISSN: 0004-5411. DOI: 10.1145/2559903.

[Yao82] Andrew C. Yao. “Theory and Application of Trapdoor Functions”. In: *Proceedings of the 23rd Symposium on Foundations of Computer Science (FOCS)*. 1982, pp. 80–91. DOI: 10.1109/SFCS.1982.45.

[Yao85] Andrew Chi-Chih Yao. “Separating the polynomial-time hierarchy by oracles”. In: *26th Symposium on Foundations of Computer Science (FOCS)*. 1985, pp. 1–10. DOI: 10.1109/SFCS.1985.49.

[Yao90] Andrew Chi-Chih Yao. “On ACC and threshold circuits”. In: *Proceedings of the 31st Symposium on Foundations of Computer Science (FOCS)*. 1990, pp. 619–627. DOI: 10.1109/FSCS.1990.89583.

A Tight Correlation Bounds for Majority

In this section, we present tight bounds on the correlation between the majority function and AC^0_d or $\mathsf{AC}^0_d[\oplus]$ circuits.

A.1 Hardness of Majority

We begin by proving Theorem 3, which says that if g is an $\mathsf{AC}^0_d[\oplus]$ circuit of size $S \geq n$, then $\mathsf{Corr}(g, \mathsf{MAJ}_n) \leq O(\log S)^{d-1}/\sqrt{n}$. A slightly weaker bound of $O(\log S)^d/\sqrt{n}$ is a known consequence of the standard Razborov-Smolensky method [Fil10]. O'Donnell and Wimmer proved the stronger bound $O(\log S)^{d-1}/\sqrt{n}$ for the special case of AC^0_d circuits [OW07], and Tal presented another proof for the special case of AC^0_d circuits where S is not too large [Tal17].

Our proof is a slight variation on the known Razborov-Smolensky argument. We rely on standard probabilistic $\mathbb{F}_2$-polynomials for the AND and OR functions.

Lemma 9 (Probabilistic polynomials for AND and OR [Raz87])**.** *For every $n \in \mathbb{N}$ and $\varepsilon > 0$, there exists a distribution over polynomials $\mathbf{p}\colon \mathbb{F}_2^n \to \mathbb{F}_2$ of degree at most $O(\log(1/\varepsilon))$ such that for every $x \in \mathbb{F}_2^n$,*

$$\Pr_{\mathbf{p}}[\mathbf{p}(x) = \mathsf{AND}(x_1, \dots, x_n)] \geq 1 - \varepsilon.$$

Similarly, there exists a distribution over polynomials $\mathbf{p}\colon \mathbb{F}_2^n \to \mathbb{F}_2$ of degree at most $O(\log(1/\varepsilon))$ such that for every $x \in \mathbb{F}_2^n$,

$$\Pr_{\mathbf{p}}[\mathbf{p}(x) = \mathsf{OR}(x_1, \dots, x_n)] \geq 1 - \varepsilon.$$

We also rely on the known bound on the correlation between the majority function and low-degree $\mathbb{F}_2$-polynomials.

Lemma 10 (Optimal bound on correlation between majority and low-degree polynomials over $\mathbb{F}_2$ [Smo87; Sze89; Smo93; Vio19])**.** *If $p\colon \mathbb{F}_2^n \to \mathbb{F}_2$ is a polynomial of degree at most k, then*

$$\Pr_{\mathbf{x} \in \mathbb{F}_2^n}[p(\mathbf{x}) = \mathsf{MAJ}_n(\mathbf{x})] \leq \frac{1}{2} + O(k/\sqrt{n}).$$

Finally, we rely on standard bounds on the Fourier coefficients of the AND and OR functions.

Lemma 11 (Fourier L_1 bounds for AND and OR)**.** *Let $f\colon \{\pm 1\}^n \to \{\pm 1\}$ be either the AND function (i.e., MAX) or the OR function (i.e., MIN). Then it is possible to write f in the form*

$$f(x) = \sum_{S \subseteq [n]} c_S \cdot \prod_{i \in S} x_i$$

(a multilinear polynomial over $\mathbb{R}$) where $\sum_{S \subseteq [n]} |c_S| \leq O(1)$.

Proof of Theorem 3. First, assume that the output gate of g is a parity gate. For this first part, it is most convenient to think of all wires as carrying $\{0, 1\}$ values, so g is a function $g\colon \mathbb{F}_2^n \to \mathbb{F}_2$. For each gate ϕ of g *other than* the output gate independently, sample a probabilistic polynomial $\mathbf{p}_\phi$ that simulates ϕ via Lemma 9 with error $\varepsilon = 1/(Sn)$. Since parity is sum mod two, the output gate can be computed exactly by a polynomial of degree 1. Therefore, by composing these polynomials, we get a random polynomial $\mathbf{p}\colon \mathbb{F}_2^n \to \mathbb{F}_2$ such that $\Pr_{\mathbf{p}}[\mathbf{p}(x) = g(x)] \geq 1 - 1/n$, and the degree of $\mathbf{p}$ is at most $O(\log S)^{d-1}$. Therefore,

$$\Pr_{\mathbf{x} \in \mathbb{F}_2^n}[g(\mathbf{x}) = \mathsf{MAJ}_n(\mathbf{x})] \leq \Pr_{\mathbf{x},\mathbf{p}}[g(\mathbf{x}) \neq \mathbf{p}(\mathbf{x})] + \Pr_{\mathbf{x},\mathbf{p}}[\mathbf{p}(\mathbf{x}) = \mathsf{MAJ}_n(\mathbf{x})] \leq \frac{1}{2} + O(\log S)^{d-1}/\sqrt{n} + 1/n,$$

completing the proof in this case.

Next, suppose the output gate is AND or OR. For this part, it is most convenient to think of all wires as carrying $\{\pm 1\}$ values, so g is a function $g\colon \{\pm 1\}^n \to \{\pm 1\}$. By Lemma 11, we can write $g = \sum_i c_i g_i$, where each g_i is an $\mathsf{AC}^0_d[\oplus]$ circuit of size S with a parity gate on top (note that the product of $\{\pm 1\}$ values is the parity of the corresponding $\{0, 1\}$ values) and $\sum_i |c_i| \leq O(1)$. Therefore,

$$\mathsf{Corr}(g, \mathsf{MAJ}_n) = \sum_i c_i \cdot \mathsf{Corr}(g_i, \mathsf{MAJ}_n) \leq \sum_i |c_i| \cdot O(\log S)^{d-1}/\sqrt{n} \leq O(\log S)^{d-1}/\sqrt{n}. \qquad \square$$

A.2 Tightness: AC^0 Circuits That Correlate with Majority

We now show that the correlation bound of Theorem 3 is tight. The proof is based on a construction due to Rossman and Srinivasan [RS19], building on prior work by O'Donnell and Wimmer [OW07] and Amano [Ama09]. Rossman and Srinivasan constructed an AC_d^0 circuit for solving the so-called "coin problem" with the following parameters.

Theorem 13 (AC_d^0 circuits solving the coin problem [RS19])**.** *For every $\gamma, \delta > 0$ and $d \in \mathbb{N}$ such that $2 \le d \le \frac{\alpha \log(\delta/\gamma)}{\log\log(\delta/\gamma)}$ for a suitable constant $\alpha > 0$, there exists a monotone AC_d^0 circuit $g\colon \{\pm1\}^r \to \{\pm1\}$ of size*

$$\exp\left(\left(\frac{1}{\gamma}\right)^{1/(d-1)} \cdot O\left(\frac{1}{\delta}\right)^{1-1/(d-1)} \cdot \log(1/\delta)\right)$$

such that if $\mathbf{x}_1, \dots, \mathbf{x}_r \in \{\pm1\}$ are independent and identically distributed bits with $\Pr[\mathbf{x}_i = +1] = 1/2 + \gamma$ for every i, then for every $b \in \{\pm1\}$, we have

$$\Pr[g(b \cdot \mathbf{x}) = b] \ge 1 - \delta.$$

Prior work starting with O'Donnell and Wimmer [OW07] has shown that monotone circuits solving the coin problem imply circuits that compute the majority function on a *large* $(1-\delta)$-fraction of inputs; indeed, Rossman and Srinivasan stated their result in terms of the latter problem [RS19]. We now show by a similar argument that monotone circuits solving the coin problem imply relatively *small* circuits that have nontrivial *correlation* with the majority function, i.e., they successfully compute the majority function on slightly more than half of inputs. Since the vanishing-failure-probability case is covered by Rossman and Srinivasan's work [RS19], the theorem below focuses only on the case that the success probability is at most $3/4$.

Theorem 14 (AC_d^0 circuits for majority with optimal correlation)**.** *For every $n, d, S \in \mathbb{N}$ where $d \le \frac{\alpha \log n}{\log\log n}$ for a suitable constant $\alpha > 0$, there exists an AC_d^0 circuit $g\colon \{\pm1\}^n \to \{\pm1\}$ of size at most S such that*

$$\mathsf{Corr}(g, \mathsf{MAJ}_n) \ge \min\left\{\frac{\Omega(\log S)^{d-1}}{\sqrt{n}}, \frac{1}{2}\right\}.$$

To prove Theorem 14, we rely on the following estimate for near-central binomial coefficients.

Proposition 9 (Lower bound on near-central binomial coefficients)**.** *Let n, Δ be positive integers such that n is even and $\Delta \le n/4$. Then*

$$\binom{n}{n/2+\Delta} \ge \Omega\left(\frac{2^n}{\sqrt{n} \cdot \exp(4\Delta^2/n)}\right).$$

Proof. It is well-known that $\binom{n}{n/2} = \Theta(2^n/\sqrt{n})$. Furthermore,

$$\begin{aligned}
\frac{\binom{n}{n/2}}{\binom{n}{n/2+\Delta}} = \frac{(n/2+\Delta)! \cdot (n/2-\Delta)!}{(n/2)! \cdot (n/2)!} &= \frac{n/2+\Delta}{n/2} \cdot \frac{n/2+\Delta-1}{n/2-1} \cdots \frac{n/2+1}{n/2-\Delta+1} \\
&= \left(1 + \frac{\Delta}{n/2}\right) \cdot \left(1 + \frac{\Delta}{n/2-1}\right) \cdots \left(1 + \frac{\Delta}{n/2-\Delta+1}\right) \\
&\le \left(1 + \frac{4\Delta}{n}\right)^{\Delta} \\
&\le \exp(4\Delta^2/n).
\end{aligned}$$

□

Proof of Theorem 14. If $d = 1$, then the theorem can be proven by taking $g(x) = x_1$, so assume $d \geq 2$. Let δ be a small enough constant, and let $\gamma = 1/\Theta(\log S)^{d-1}$ be the smallest value such that the size bound $\exp(O((1/\gamma)^{1/(d-1)}))$ in Theorem 13 is at most S. Let $g\colon \{\pm 1\}^r \to \{\pm 1\}$ be the corresponding circuit of size S. If $\delta^2/\gamma^2 \leq n/2$, then let m be the largest positive integer such that $m \leq \delta^2/\gamma^2$ and $n - m$ is even; otherwise let $m = n$. Sample a list of indices $\mathbf{i} \in [m]^r$ uniformly at random. Let $\mathbf{g}\colon \{\pm 1\}^m \times \{\pm 1\}^{n-m} \to \{\pm 1\}$ be the random circuit defined by

$$\mathbf{g}(x, y) = g(x_{\mathbf{i}}) = g(x_{\mathbf{i}_1}, \ldots, x_{\mathbf{i}_r}).$$

We use the notation $\Sigma(x) = \sum_i x_i$ introduced in Section 5. We can write the correlation between $\mathbf{g}$ and MAJ_n as follows.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y},\mathbf{g}}[\mathbf{g}(\mathbf{x}, \mathbf{y}) \cdot \mathsf{MAJ}_n(\mathbf{x}, \mathbf{y})] \\
&\quad = \sum_{\Sigma=-m}^{m} \Pr_{\mathbf{x}}[\Sigma(\mathbf{x}) = \Sigma] \cdot \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{i},\mathbf{y}}[g(\mathbf{x}_{\mathbf{i}}) \cdot \mathsf{MAJ}_n(\mathbf{x}, \mathbf{y}) \mid \Sigma(\mathbf{x}) = \Sigma] \\
&\quad = \sum_{\Sigma=-m}^{m} \Pr_{\mathbf{x}}[\Sigma(\mathbf{x}) = \Sigma] \cdot \underbrace{\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{i}}[g(\mathbf{x}_{\mathbf{i}}) \cdot \operatorname{sign}(\Sigma) \mid \Sigma(\mathbf{x}) = \Sigma]}_{(*)} \cdot \underbrace{\mathop{\mathbb{E}}_{\mathbf{y}}[\operatorname{sign}(\Sigma + \Sigma(\mathbf{y})) \cdot \operatorname{sign}(\Sigma)]}_{(**)}.
\end{aligned}$$

Let us consider a fixed $\Sigma \in \{-m, -m+1, \ldots, m\}$. Let $\Sigma_* = \lceil 2\gamma m \rceil$, and let $\varepsilon_* = \mathbb{E}_{\mathbf{y}}[\operatorname{sign}(\Sigma_* + \Sigma(\mathbf{y}))]$. If $|\Sigma| \geq \Sigma_*$, then the quantity (*) is at least $1 - 2\delta$ by the correctness of g and the quantity (**) is at least ε_*. Meanwhile, if $|\Sigma| < \Sigma_*$, then quantity (*) is at least -1 and quantity (**) is at most ε_*. Therefore, we get

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y},\mathbf{g}}[\mathbf{g}(\mathbf{x}, \mathbf{y}) \cdot \mathsf{MAJ}_n(\mathbf{x}, \mathbf{y})] \geq (1 - 2\delta) \cdot \varepsilon_* \cdot \Pr_{\mathbf{x}}[|\Sigma(\mathbf{x})| \geq \Sigma_*] - \varepsilon_* \cdot \Pr_{\mathbf{x}}[|\Sigma(\mathbf{x})| < \Sigma_*].$$

We have $\Pr_{\mathbf{x}}[|\Sigma(\mathbf{x})| < \Sigma_*] \leq O(\frac{\Sigma_*}{\sqrt{m}}) = O(\delta)$ because every binomial coefficient $\binom{m}{k}$ is at most $O(2^m/\sqrt{m})$. Therefore,

$$\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y},\mathbf{g}}[\mathbf{g}(\mathbf{x}, \mathbf{y}) \cdot \mathsf{MAJ}_n(\mathbf{x}, \mathbf{y})] \geq (1 - 2\delta) \cdot \varepsilon_* \cdot (1 - O(\delta)) - \varepsilon_* \cdot O(\delta) > \varepsilon_* \cdot (1 - O(\delta)).$$

The best case is at least as good as the average case, so there is some fixing g of $\mathbf{g}$ such that

$$\mathsf{Corr}(g, \mathsf{MAJ}_n) \geq \varepsilon_* \cdot (1 - O(\delta)).$$

Observe that

$$\varepsilon_* = \Pr_{\mathbf{y}}[|\Sigma(\mathbf{y})| \leq \Sigma_*].$$

Now we split into two cases. First, suppose $\Sigma_* \leq \sqrt{2\ln(1/\delta) \cdot (n - m)}$. Then $n - m \geq n/2$, so (assuming n is sufficiently large) we have $\Sigma_*/2 \leq (n - m)/4$, and hence we may apply Proposition 4 to get

$$\begin{aligned}
\varepsilon_* \geq \sum_{\Delta=-\lfloor \Sigma_*/2 \rfloor}^{\lfloor \Sigma_*/2 \rfloor} \binom{n-m}{(n-m)/2 + \Delta} \cdot 2^{-(n-m)} \geq \frac{\Sigma_*}{\sqrt{n - m} \cdot \exp(O(\ln(1/\delta)))} &\geq \frac{\gamma \cdot \delta^{O(1)}}{\sqrt{n}} \\
&= \frac{\Omega(\log S)^{d-1} \cdot \delta^{O(1)}}{\sqrt{n}}.
\end{aligned}$$

On the other hand, if $\Sigma_* > \sqrt{2\ln(1/\delta) \cdot (n - m)}$, then Hoeffding's inequality gives $\varepsilon_* \geq 1 - 2\delta$, and hence we get correlation at least $1/2$ provided we choose δ to be a small enough constant. □

B Near-Tight Bounds for Approximations by Depth-One Circuits

In this section, we show that prior work readily implies nearly-matching upper and lower bounds regarding the task of approximating an AC^0_{1+k} circuit by an AC^0_1 circuit. We begin with the construction, showing that AC^0_1 circuits *can* nontrivially approximate AC^0_{1+k} circuits. Previously, Hatami, Hoza, Tal, and Tell observed [HHTT23, Proposition A.3] that this follows from the Linial-Mansour-Nisan theorem [LMN93]. The proposition below slightly refines their argument to get a tighter bound. We assume k is a constant for simplicity.

Proposition 10 (Using AC^0_1 circuits to approximate AC^0_{1+k} circuits)**.** *Let $k \in \mathbb{N}$ be a constant. Let $h\colon \{\pm 1\}^n \to \{\pm 1\}$ be an AC^0_{1+k} circuit of size S. There exists an AC^0_1 circuit (i.e., a conjunction or disjunction of literals) g with 1 gate and $O(\log^k S)$ wires such that*

$$\mathsf{Corr}(g,h) \geq \exp\left(-O\left((\log S)^k \cdot \log\log S\right)\right).$$

Proof. We rely on bounds on the Fourier spectrum of h [LMN93; Bop97; Hås01; Tal17]. Every function computable by an AC^0 circuit is concentrated on a relatively *small collection* of Fourier coefficients. In particular, Tal showed that there is a collection $\mathcal{T}$ of subsets $T \subseteq [n]$ such that $|\mathcal{T}| \leq \exp(O((\log S)^k \cdot \log\log S))$ and $\sum_{T\in\mathcal{T}} \widehat{h}(T)^2 \geq 2/3$ [Tal17]. Furthermore, every function computable by an AC^0 circuit is concentrated on its *low-degree* Fourier coefficients. In particular, Tal showed that there is a value $\ell = O(\log^k S)$ such that $\sum_{T\subseteq[n],|T|>\ell} \widehat{h}(T)^2 \leq 1/3$ [Tal17]. Combining these two bounds, we see that $\sum_{T\in\mathcal{T},|T|\leq\ell} \widehat{h}(T)^2 \geq 1/3$. Therefore, there is some $T_* \subseteq [n]$ such that $|T_*| \leq \ell$ and $\widehat{h}(T_*)^2 \geq \frac{1}{3|\mathcal{T}|}$.

The fact that $\widehat{h}(T_*)$ is relatively large means that h is approximated reasonably well by the parity function

$$\chi_{T_*}(x) := \prod_{i\in T_*} x_i.$$

To get an AC^0_1 approximation, we now write χ_{T_*} as a linear combination of (the $\{\pm1\}$-valued versions of) conjunctions of literals. For each string $a \in \{\pm1\}^{|T_*|}$, define $\mathsf{MATCH}_a\colon \{\pm1\}^n \to \{\pm1\}$ by the rule

$$\mathsf{MATCH}_a(x) = \begin{cases} -1 & \text{if } x_{T_*} = a \\ 1 & \text{otherwise.} \end{cases}$$

Then

$$\sum_a \mathsf{MATCH}_a(x) \cdot \prod_{i\in T_*} a_i = \sum_a \prod_i a_i - 2\cdot\chi_{T_*}(x),$$

and hence

$$\chi_{T_*}(x) = -\frac{1}{2}\cdot\sum_a \mathsf{MATCH}_a(x)\cdot\prod_i a_i.$$

Therefore,

$$\widehat{h}(T_*) = \mathsf{Corr}(h,\chi_{T_*}) = -\frac{1}{2}\cdot\sum_a \mathsf{Corr}(h,\mathsf{MATCH}_a)\cdot\prod_i a_i,$$

and so there must be some $a_* \in \{\pm1\}^{|T_*|}$ such that $|\mathsf{Corr}(h,\mathsf{MATCH}_a)| \geq 2\cdot 2^{-\ell}\cdot|\widehat{h}(T_*)|$. Depending on whether $\mathsf{Corr}(h,\mathsf{MATCH}_a)$ is positive or negative, we either let $g = \mathsf{MATCH}_a$ (a conjunction of literals) or $g = -\mathsf{MATCH}_a$ (a disjunction of literals). Either way, we get

$$\mathsf{Corr}(g,h) \geq \frac{2\cdot 2^{-\ell}}{\sqrt{3|\mathcal{T}|}} = \exp\left(-O\left((\log S)^k\cdot\log\log S\right)\right). \qquad \square$$

Now we show that AC^0_1 circuits *cannot* approximate AC^0_{1+k} circuits significantly better than the construction of Proposition 10. The proof is rather trivial: the hard function is the parity function on an appropriate number of bits. Again, for simplicity, we assume k is a constant.

Proposition 11 (Hardness of approximating AC^0_{1+k} circuits using AC^0_1 circuits)**.** *Let $k \in \mathbb{N}$ be a constant. There exists an AC^0_{1+k} circuit $h \colon \{\pm 1\}^n \to \{\pm 1\}$ of size $O(n)$ such that for every AC^0_1 circuit g, we have*

$$\mathsf{Corr}(g, h) \leq \exp\left(-\Omega\left(\log^k n\right)\right).$$

Proof. Let $t = \lfloor \log^k(\sqrt{n}) \rfloor$, and let $h(x)$ be the parity of the first t bits of x. Then h can be computed by an AC^0_{1+k} circuit of size $\widetilde{O}(\sqrt{n})$ (Proposition 1). Now let g be any AC^0_1 circuit. Without loss of generality, we assume that only the first t variables appear in g, and we assume that each variable appears in g at most once. If g reads fewer than t variables, then it is easy to see that $\mathsf{Corr}(g, h) = 0$, so we may assume that each of the first t variables appears exactly once in g. Let b be the less-likely output value of g, i.e., $b = -1$ if g is a conjunction and $b = +1$ if g is a disjunction. Then

$$\begin{aligned}
\Pr[g(\mathbf{x}) = h(\mathbf{x})] &= \Pr[g(\mathbf{x}) = h(\mathbf{x}) = +1] + \Pr[g(\mathbf{x}) = h(\mathbf{x}) = -1] \\
&\leq \Pr[h(\mathbf{x}) = -b] + \Pr[g(\mathbf{x}) = b] \\
&= \frac{1}{2} + 2^{-t}.
\end{aligned}$$

□